



INSTITUTE FOR DEFENSE ANALYSES

**Application of User-Oriented MOE to
Transport and Dispersion Model
Predictions of the European
Tracer Experiment**

Steve Warner, Project Leader
Nathan Platt
James F. Heagy

November 2003

Approved for public release;
distribution unlimited.

IDA Paper P-3829

Log: H 03-002088

**This work was conducted under contracts DASW01 98 C 0067/
DASW01 02 C 0012, Task DC-9-1797, for the Defense Threat Reduction
Agency. The publication of this IDA document does not indicate
endorsement by the Department of Defense, nor should the contents be
construed as reflecting the official position of that Agency.**

**© 2003 Institute for Defense Analyses, 4850 Mark Center Drive,
Alexandria, Virginia 22311-1882 • (703) 845-2000.**

**This material may be reproduced by or for the U.S. Government pursuant
to the copyright license under the clause at DFARS 252.227-7013
(NOV 95).**

INSTITUTE FOR DEFENSE ANALYSES

IDA Paper P-3829

**Application of User-Oriented MOE to
Transport and Dispersion Model
Predictions of the European
Tracer Experiment**

Steve Warner, Project Leader
Nathan Platt
James F. Heagy

PREFACE

This paper was prepared by the Institute for Defense Analyses (IDA) for the Defense Threat Reduction Agency (DTRA), in partial fulfillment of the task “Support for DTRA in the Validation Analysis of Hazardous Material Transport and Dispersion Prediction Models.” The objective of this effort was to conduct analyses and special studies associated with the verification, validation, and accreditation (VV&A) of hazardous transport and dispersion prediction models.

This paper represents the first in a planned series of three papers that compares the predictions of several transport and dispersion models to the data collected during the *European Tracer Experiment (ETEX)* release of October 1994. This first paper focuses on the methodology of comparison – that is, the previously described Measure of Effectiveness for transport and dispersion models.

The IDA Technical Review Committee was chaired by Robert R. Soule and consisted of Arthur Fries, Nelson S. Pacheco, Janet M. Pavelich, and Edward T. Toton. The authors thank Stefano Galmarini (Joint Research Centre – Environment Institute, Environment Monitoring Unit, Ispra, Italy) for both providing access to the model predictions of the *ETEX* release and for numerous useful discussions.

APPLICATION OF USER-ORIENTED MOE TO TRANSPORT AND DISPERSION MODEL PREDICITONS OF THE *EUROPEAN TRACER EXPERIMENT*

TABLE OF CONTENTS

SUMMARY	1
A. Introduction	1
B. Purpose	3
C. Results and Discussion: Model Comparisons to <i>ETEX</i>	3
D. Outline of This Paper.....	6
 1. INTRODUCTION.....	 1-1
A. Background.....	1-1
B. User-Oriented Measure of Effectiveness (MOE)	1-1
1. Computation of MOE	1-4
a. MOE Based on Concentration or Dosage.....	1-5
b. MOE Based on a Concentration or Dosage Threshold.....	1-5
c. Sensitivity of MOE Values To Specific Sampler Location.....	1-7
2. Scoring Functions for the MOE	1-7
a. Objective Scoring Function (OSF): Value Closest to 1,1	1-7
b. Risk-Weighted Figure of Merit in Space (RWFMS).....	1-8
c. Fractional Bias Figure of Merit (FB-FOM).....	1-10
d. Normalized Absolute Difference (NAD)	1-14
C. Brief Description of the <i>European Tracer Experiment (ETEX)</i>	1-15
D. Outline of this Study.....	1-21
References.....	1-R-1
 2. RESULTS AND DISCUSSION	 2-1
A. MOE Values for Predictions of <i>ETEX</i>	2-1
1. Threshold-Based MOE Values: 3-Hour Average Concentration ...	2-1
2. Summed Concentration-Based MOEs	2-10
B. Sensitivity of MOE Values To Single Sampler Locations	2-12
C. MOE Values as a Function of Time	2-18
D. Planned Future Studies Involving the MOE and <i>ETEX</i>	2-27
References.....	2-R-1

Appendix A – Acronyms

Appendix B – Task Order Extract

LIST OF FIGURES

1-1.	Conceptual View of Overlap (A_{OV}), False Negative (A_{FN}), and False Positive (A_{FP}) Regions That are Used to Construct the User-Oriented MOE.....	1-2
1-2.	Key Characteristics of the Two-Dimensional MOE Space	1-3
1-3.	Interpretation of Comparisons: Exclusionary Zones	1-4
1-4.	Illustration of MOE Computations Based for Model 121 (SCIPUFF) Predictions of the 3-Hour Average Concentrations for the Time Period Between 21 and 24 Hours After the <i>ETEX</i> Release.....	1-6
1-5.	Relationship Between RWFMS and 2D MOE, Isolines of RWFMS in the MOE Space: a) $C_{FN} = 1$, $C_{FP} = 1$, b) $C_{FN} = 5$, $C_{FP} = 0.5$	1-9
1-6.	User Coloring of MOE Space: RWFMS	1-10
1-7.	FB Isolines for Some Values of the Parameter s	1-13
1-8.	Relationship Between FB and MOE: Examples of FBFOM User-Coloring for $s = 1.15$, 1.5 and 2	1-13
1-9.	Relationship Between NAD and MOE: Isolines of NAD in the 2D MOE Space.....	1-15
1-10.	<i>ETEX</i> Sampler Locations Across Europe: Red Open Triangles Correspond to Sampler Locations and Black Open Circle Corresponds to the Release Location	1-16
1-11.	Observed PMCH Concentrations Across Europe.....	1-18
2-1.	3-Hour Average Concentration Threshold-Based MOE (0.01 ng m^{-3}) Values for 46 ATEMS II Participants.....	2-2
2-2.	3-Hour Average Concentration Threshold-Based MOE (0.1 ng m^{-3}) Values for 46 ATEMS II Participants.....	2-5
2-3.	3-Hour Average Concentration Threshold-Based MOE (0.5 ng m^{-3}) Values for 46 ATEMS II Participants.....	2-8
2-4.	MOE Values for 46 ATEMS II Participants and Based on 3-Hour Average Concentrations Comparisons	2-11
2-5.	MOE Values for SCIPUFF (Model 121) Predictions of <i>ETEX</i> and Based on a 0.10 ng m^{-3} Threshold for 3-Hour Average Concentrations Comparisons	2-14
2-6.	Summed Concentration-Based MOE Values for Six Model Predictions of <i>ETEX</i> for 3-Hour Average Concentrations Comparisons.....	2-14
2-7.	Summed Concentration-Based MOE Values for the Six Model Predictions of <i>ETEX</i> for 3-Hour Average Concentrations Comparisons That Were Most Influenced by a Single Sampler Location.....	2-15
2-8.	Sampler F21 (Rennes) and Release Point (Monterfil) Locations	2-16
2-9.	MOE Values Based on a Threshold Concentration of 0.01 ng m^{-3} for Model Predictions of <i>ETEX</i>	2-19
2-10.	MOE Values Based on a Threshold Concentration of 0.1 ng m^{-3} for Model Predictions of <i>ETEX</i>	2-21
2-11.	12-Hour RTW MOE Values Based on a Threshold Concentration of 0.01 ng m^{-3} for Six Model Predictions of <i>ETEX</i>	2-22
2-12.	12-Hour RTW MOE Values Based on a Threshold Concentration of 0.1 ng m^{-3} for Six Model Predictions of <i>ETEX</i>	2-24

2-13.	12-Hour and 24-Hour RTW MOE Values Based on a Threshold Concentration of 0.01 ng m ⁻³ for Two Model Predictions of <i>ETEX</i>	2-25
2-14.	12-Hour and 24-Hour RTW MOE Values Based on a Threshold Concentration of 0.1 ng m ⁻³ for Two “Highly-Ranked” Model Predictions of <i>ETEX</i>	2-26
2-15.	24-Hour RTW Summed Concentration MOE Values for the Top Two ABS(FB) Ranked Model Predictions of <i>ETEX</i>	2-27

LIST OF TABLES

1.	Top-Ranked Model and Rankings of SCIPUFF and ARAC Based on MOE Values and the Objective Scoring Function.....	4
2.	SCIPUFF and ARAC Rankings Based on OSF <i>or</i> NAD and the Inclusion/Exclusion of Sampler Location F21	5
1-1.	ATEMS II Participants For Which IDA Obtained Predictions	1-19
1-1.	(Cont’d).....	1-20
2-1.	Relative Model Rankings Based on 0.01 ng m ⁻³ Threshold MOE Values for Three Scoring Functions: OSF, RWFMS (1,1) and RWFMS (5,0.5).....	2-4
2-2.	Relative Model Rankings Based on 0.1 ng m ⁻³ Threshold MOE Values for Three Scoring Functions: OSF, RWFMS (1,1) and RWFMS (5,0.5).....	2-7
2-3.	Relative Model Rankings Based on 0.5 ng m ⁻³ Threshold MOE Values for Three Scoring Functions: OSF, RWFMS (1,1) and RWFMS (5,0.5).....	2-9
2-4.	Relative Model Rankings Based on Summed Concentration MOE Values for Five Scoring Functions: OSF, RWFMS (1,1), NAD, RWFMS (5,0.5), and FB (i.e., absolute value of FB)	2-12
2-5.	Relative Model Rankings Based on Summed Concentration MOE Values for Five Scoring Functions After the Removal of Sampler Location F21: OSF, RWFMS (1,1), NAD, RWFMS (5,0.5), and FB.....	2-17

SUMMARY

A. INTRODUCTION

In October 1994, the inert, environmentally safe, tracer gas perfluoro-methyl-cyclohexane (PMCH) was released over a 12-hour period from a location in northwestern France and tracked at 168 sampling locations in 17 countries across Europe (hundreds of kilometers).¹ This release, known as the European Tracer Experiment (*ETEX*), resulted in the collection of a wealth of data. IDA has obtained from the Joint Research Centre, European Commission 46 sets of transport and dispersion predictions associated with models from 17 countries (Table 1-1) – including HPAC/SCIPIUFF and ARAC (LLNL)² – as well as the observed PMCH sampling data associated with the October 1994 *ETEX* release.³ This paper describes the extension of the previously developed user-oriented two-dimensional measure of effectiveness (MOE) methodology to evaluate the predictions of these 46 models against the long-range *ETEX* observations.

The two-dimensional MOE allows for the evaluation of transport and dispersion model predictions in terms of “false negative” (under-prediction) and “false positive” (over-prediction) regions.⁴ A perfect model prediction leads to no false negative *and* no

¹ Graziani, G., Klug, W., and Mosca, S., 1998: *Real-Time Long-Range Dispersion Model Evaluation of the ETEX First Release*, Joint Research Center, European Commission, Office of Official Publications of the European communities, L-2985 (CL-NA-17754-EN-C), Luxembourg, 1998.

² HPAC = Hazardous Prediction and Assessment Capability, SCIPIUFF = Second-Order Closure Integrated Puff, ARAC = Atmospheric Release Advisory Center, and LLNL = Lawrence Livermore National Laboratory. Since HPAC/SCIPIUFF and ARAC (now known as NARAC – National ARAC), are of particular interest to our sponsor, we typically focus extra attention in the paper on the results associated with the predictions of these two models.

³ Mosca, S., Bianconi, R., Bellasio, R., Graziani, G., and Klug, W., 1998: *ATEMS II – Evaluation of Long-Range Dispersion Models Using Data of the 1st ETEX Release*, Joint Research Center, European Commission, Office of Official Publications of the European communities, L-2985 (CL-NA-17756-EN-C), Luxembourg, 1998.

⁴ Warner, S., Platt, N., and Heagy, J. F., 2004: “User-Oriented Two-Dimensional Measure of Effectiveness for the Evaluation of Transport and Dispersion Models,” in press *J. Appl. Meteor.* and Warner S., Platt, N., and Heagy, 2001: “User-Oriented Measures of Effectiveness for the Evaluation of Transport and Dispersion Models,” *Proceedings of the Seventh International Conference on Harmonisation Within Atmospheric Dispersion Modelling for Regulatory Purposes*, Belgirate, Italy, 28-21 May 2001, pages 24-29.

false positive, that is, complete and perfect overlap of the predictions and observations. Such a perfect model would have a two-dimensional MOE value of (1,1).⁵ For a given application and user risk tolerance, certain regions of the two-dimensional MOE space may be considered acceptable. For example, some users may tolerate a certain false positive fraction (ultimately, unnecessarily warned individuals) but require a very low false negative fraction (inadvertently exposed individuals). Such a risk tolerance profile implies a certain location in the two-dimensional MOE space (see Chapter 1) and can be turned into a mathematical function for “scoring” the MOE predictions. Other user “scoring” functions also have been developed for the MOE.⁶

MOE values can be computed by considering the prediction of concentrations summed across all sampler locations or MOE values can be computed based on defining a critical threshold. For threshold-based MOE values, the model is judged by its ability to predict which locations led to observations above a certain specified threshold. We calculated threshold-based MOE values for three thresholds: 0.01⁷, 0.1, and 0.5 ng m⁻³.

Our current research associated with *ETEX* is divided into three phases. First, methodological protocols have been developed to compare model predictions of *ETEX* using the MOE and to “score and rank” model performance by a variety of notional user criteria. Next, the sensitivity of MOE estimates, and hence model rankings, to any single sampler location, has been explored. The second phase of research (currently ongoing) considers converting nominal MOE estimates into true area-based MOE values. For this research, extensive analysis of interpolation schemes is being conducted and possible sensitivities are being explored. With the application of actual European population density distributions and the consideration of a notional hazardous agent, one can extend this work to describe transport and dispersion model performance in terms of falsely warned populations and inadvertently exposed populations. This is the ultimate goal of the second phase of this *ETEX* effort. The final phase of this research entails redoing HPAC/SCIPUFF predictions of *ETEX* (with the latest version of the HPAC software) to

⁵ A model prediction that completely misses the observation (perhaps, the “plume” goes in the exact opposite direction) would achieve an MOE value of (0,0).

⁶ Warner, S., Platt, N., and Heagy, J. F., 2001: *Application of User-Oriented MOE to HPAC Probabilistic Predictions of Prairie Grass Field*, IDA Paper P-3586, 275 pp, May 2001. (Available electronically [DTIC STINET ada391653] or on CD via an e-mail request to Steve Warner at swarner@ida.org or a mail request to Steve Warner, Institute for Defense Analyses, 4850 Mark Center Drive, Alexandria, Virginia 22311-1882.)

⁷ The value 0.01 ng m⁻³ was considered a lower bound by the experimenters. The experimenters treated any measurement below 0.01 ng m⁻³ as a zero (see page 11 of the reference cited in footnote 1).

include HPAC probabilistic predictions and applying the techniques developed in the first two research phases to evaluate and compare these new predictions of *ETEX*.

This paper describes the techniques and results associated with the first phase of these *ETEX* studies.

B. PURPOSE

The purpose of this paper is to extend the application of the user-oriented two-dimensional MOE to the evaluation of predictions of very long-range (hundreds of kilometers) transport and dispersion. In doing so, two objectives are to be achieved. First, estimates of transport and dispersion model MOE values for both the prediction of summed concentrations and the prediction of exceedance of specified concentration thresholds are developed in this paper. These values can serve as a *baseline* for future transport and dispersion model prediction comparisons to *ETEX*. Next, this paper describes methodological procedures that will serve as the basis for future analyses of transport and dispersion model predictions of *ETEX*.

C. RESULTS AND DISCUSSION: MODEL COMPARISONS TO *ETEX*

Forty-six sets of transport and dispersion model predictions of *ETEX* observed concentrations (3-hour average) are evaluated with the MOE in this paper. Model predictions are ranked by which model achieves the best – closest to (1,1) – MOE value (Chapter 2). We refer to this “closest to (1,1) scoring” as the objective scoring function (OSF). In addition, mathematical relationships between the MOE and a measure of bias (fractional bias – FB), a measure of scatter between observations and predictions (normalized absolute difference – NAD), and a measure that assesses spatial correlations (figure of merit in space – FMS)⁸ are described in Chapter 1.⁹ Model predictions are scored and ranked based on OSF, FB, NAD, and FMS.

Table 1 identifies the top ranked model predictions as judged by the OSF as well as the rankings (out of 46) of SCIPUFF and ARAC. Rankings are identified for the three

⁸ Mosca, S., Graziani, G., Klug, W., Bellasio, R., and Bianconi, R., 1998: “A Statistical Methodology for the Evaluation of Long-Range Dispersion Models: An Application to the *ETEX* Exercise,” *Atmos. Environ.*, 32 (24), 4307-4324.

⁹ Furthermore, a version of FMS is described in Chapter 1 that allows a user to weight the relative influence of false negative and false positive fractions on the ultimate MOE score, therefore allowing a user to impose a specified risk tolerance/aversion on the process of transport and dispersion model evaluation.

threshold-based and summed concentration-based MOE values. No single model dominated the top ranking. Complete rankings can be found in Chapter 2. Rankings based on FMS and NAD were found to be quite similar to those based on OSF.

Table 1. Top-Ranked Model and Rankings of SCIPUFF and ARAC Based on MOE Values and the Objective Scoring Function

Rank	0.01 ng m ⁻³	0.1 ng m ⁻³	0.5 ng m ⁻³	Summed Concentration
1	Canadian Meteorological Centre	Swedish Meteorological and Hydrological Office	ARAC	German Weather Service
Model	0.01 ng m ⁻³	0.1 ng m ⁻³	0.5 ng m ⁻³	Summed Concentration
SCIPUFF	24	30	23	41
ARAC	4	5	1	33

The rankings described in this paper result from consideration of a single release and *general* inference about which model is “best” or ranked highest is not appropriate. Rather, these rankings describe performance in terms of *this specific release only*. In addition, for this *single release* field experiment, no direct measures of uncertainty associated with the computed MOE values or model rankings were readily available. However, variations in MOE values as a function of time after the release and sensitivities of the MOE values and rankings to the influence of a single sampler location are briefly described below.

Past analysis¹⁰ has suggested that assessments of model performance could be sensitive to the results associated with a single sampling location – in particular, the location closest to the release where the concentrations would be highest. We examined the sensitivity of MOE values to this phenomenon by re-computing MOE values after the removal of a single sampling location. Each of the 168 sampling locations was removed (one at a time) generating 168 additional MOE values. We found that for MOE values based on summed concentrations (but not threshold exceedance), there was indeed a

¹⁰ Sykes, R. I., et al., 2000: *PC-SCIPUFF Version 1.3 Technical Documentation*, A.R.A.P Report No. 725, Titan Corporation, ARAP Group, December 2000, pages 221- 226.

sensitivity associated with the sampling location closest to the release – at Rennes, France and designated “F21.” While most of the models’ MOE values were relatively unaffected by the removal of F21, a few were, perhaps overly influenced by this single sampler location. The two models of particular interest here, DTRA’s SCIPUFF and LLNL’s ARAC, were two of about 8 (of 46) that resulted in MOE values that were significantly influenced by the removal of the single sampler location at F21. Table 2 provides the OSF-based and NAD-based model rankings (out of 46) for SCIPUFF and ARAC based on the inclusion of all 168 sampler locations and based on the exclusion of the single sampler location at F21. The rankings for SCIPUFF and ARAC were identical for both scoring functions – OSF and NAD. Other scoring functions led to similar and consistent findings given the removal of F21.

Table 2. SCIPUFF and ARAC Rankings Based on OSF or NAD and the Inclusion/Exclusion of Sampler Location F21

Model	All 168 Sampler Locations	Minus F21 (Rennes, France)
SCIPUFF	41	34
ARAC	33	8

Finally, this paper presents analysis of the variation in model predictive performance, as judged by the MOE, as a function of time. Portions of the *ETEX* sampling network were monitored out to 90 hours after the release. We compared 3-hour average concentrations (predictions and observations) for 30 time periods and also examined 12-hour running time window (i.e., 4 time periods in sequence combined) and 24-hour running time window (i.e., 8 time periods in sequence combined) MOE values. When judging model predictive performance using the MOE based on the 0.01 or 0.1 ng m⁻³ threshold, one of two time-dependent behaviors was typically observed. For some models, an initial under-prediction of the number of locations that exceed the threshold is followed by a “correction” that leads to about the right number of locations predicted above the threshold, followed finally, by degradation that suggests a general missing of the locations at which the threshold is exceeded at the longest times (and distances). For other models, an initial over-prediction of the number of locations that exceed the threshold is followed by a “correction” that leads to about the right number of locations predicted above the threshold, followed again, by degradation that suggests a general missing of the locations at which the threshold is exceeded. SCIPUFF and ARAC both show this degradation (as judged by the 0.01 or 0.1 ng m⁻³ threshold-

based MOE) at the longest times after the release, as do most of the examined transport and dispersion models.

D. OUTLINE OF THIS PAPER

This paper is divided into two chapters. Chapter 1 describes the user-oriented two-dimensional MOE and develops notional scoring functions that can be used to evaluate model predictive performance within the context of a specified user need. Brief descriptions of the *ETEX* release and the models included in this study are also provided in Chapter 1. The results of this analysis, along with some discussion, are presented in Chapter 2. Appendix A provides a list on acronyms and Appendix B provides an extract from the task order that supported this research.

CHAPTER 1
INTRODUCTION

1. INTRODUCTION.

A. BACKGROUND

In general, model validation efforts include specific measures of effectiveness (MOEs) that are needed to define a metric by which field trial observations and predictions can be compared. It is helpful if model validation includes an MOE that relates “operational” use of the model to field trial experiments. Such an MOE gives a certain degree of confidence to users with respect to how closely the model approximates the real world in their particular situation.

Previously, we developed and described a user-oriented MOE [Refs. 1-1 and 1-2]. This two-dimensional (2D) MOE has been applied to short-range [Ref. 1-3] and mid-range [Ref. 1-4] field observations, as well as predictions of an interior building release [Ref. 1-5]. Also, this 2D MOE has been used as a diagnostic aid for examining differences between sets of model predictions of field observations [Ref. 1-6] and of computer-simulated releases [Ref. 1-7]. In addition, 2D MOE values have been used to explore the differences between HPAC probabilistic outputs of short-range field observations [Ref. 1-8]. Most recently, this methodology has been applied to examine predictions of transport and dispersion in an urban environment [Ref. 1-9] and to study short-range predictions of dispersal from an improvised radiological dispersion device [Ref. 1-10].

This paper extends the application of the MOE to long-range field observations, namely the first *European Tracer Experiment (ETEX)* release of October 1994 that tracked material for 90 hours and thousands of miles [Refs. 1-11 through 1-13]. In particular, this paper provides MOE values for 46 sets of *ETEX* predictions.

B. USER-ORIENTED MEASURE OF EFFECTIVENESS (MOE)

A fundamental feature of any comparison of hazard prediction model output to observations is the over- and under-prediction regions. We define the false negative region where a hazard is observed but not predicted, and the false positive region where a hazard is predicted but not observed. Figure 1-1 shows one possible interpretation of these regions – the observed and predicted areas in which a prescribed dosage is

exceeded. This view can be extended to consider the marginal over- and under-predicted values as will be discussed below. In any case, numerical estimates of the false negative region (A_{FN}), the false positive region (A_{FP}), and the overlap region (A_{OV}) characterize this conceptual view.

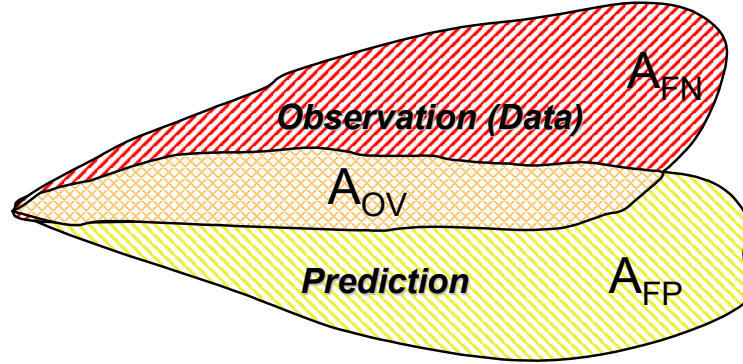


Figure 1-1. Conceptual View of Overlap (A_{OV}), False Negative (A_{FN}), and False Positive (A_{FP}) Regions That are Used to Construct the User-Oriented MOE

The MOE that we consider has two dimensions. The x-axis corresponds to the ratio of overlap region to the observed region and the y-axis corresponds to the ratio of overlap region to the predicted region. When these mathematical definitions are algebraically rearranged (Eq. 1-1 below), we recognize that the x-axis corresponds to *1 minus the false negative fraction* and the y-axis corresponds to *1 minus the false positive fraction*,

$$MOE = (x, y) = \left(\frac{A_{OV}}{A_{OB}}, \frac{A_{OV}}{A_{PR}} \right) = \left(\frac{A_{OB} - A_{FN}}{A_{OB}}, \frac{A_{PR} - A_{FP}}{A_{PR}} \right) = \left(1 - \frac{A_{FN}}{A_{OB}}, 1 - \frac{A_{FP}}{A_{PR}} \right). \quad (1-1)$$

where A_{FN} = region of false negative, A_{FP} = region of false positive, A_{OV} = region of overlap, A_{PR} = region of the prediction, and A_{OB} = region of the observation. Consistent with the above algebraic rearrangement, Figure 1-2 shows the region of false negative decreasing from left to right and the region of the false positive decreasing from bottom to top.

Figure 1-2 demonstrates some of the key characteristics of the 2D MOE space. We begin with the (1,1) point located at the upper-right corner. Here, both plumes overlap entirely (no false negative nor false positive fraction), and thus the model would achieve perfect agreement with the field trial. Point (0,0) signifies that there is no region of overlap, and thus the model disagrees completely with the field trial. This 2D MOE includes directional effects; that is, the prediction of the location of a hazard and not just the shape and size of the plume is critical to obtaining a high MOE “score.”

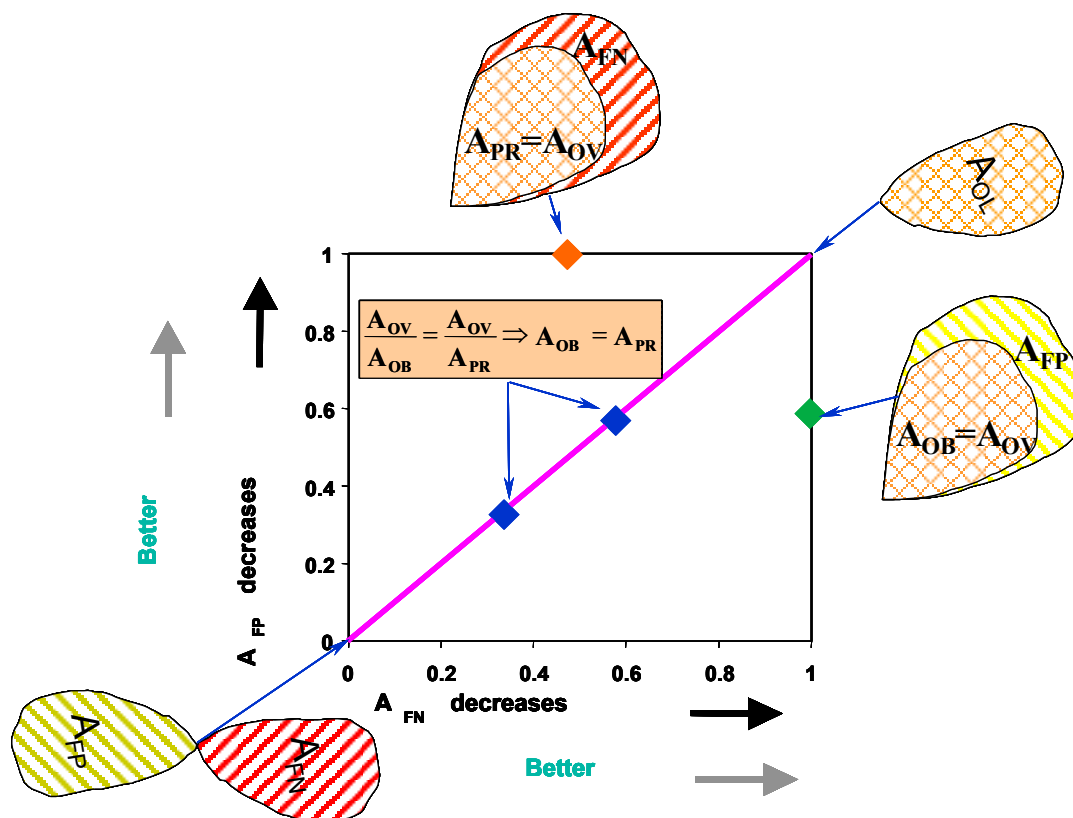


Figure 1-2. Key Characteristics of the Two-Dimensional MOE Space

Point (1,0) represents a situation where there is no false negative region, while there is an “infinite” false positive region (for nonzero releases). At this point, the implication is that the model predicts hazard everywhere. Along the line, $x = 1$, the prediction completely envelops the observation.

Point (0,1) signifies that there is no false positive region, but there is an “infinite” false negative region (for nonzero releases). At this point, the implication is that the model predicts no hazard. Along the line, $y = 1$, the observation completely envelops the prediction.

The “purple” diagonal line represents the situation where the prediction and the observation have identical “total” sizes (that is, $x = y$ implies from Eq. (1-1) that $A_{OB} = A_{PR}$). As one traverses this diagonal line from (1,1) toward (0,0), the fraction of overlap region between the predicted and observed plumes decreases.

Figure 1-3 suggests an additional interpretation of the 2D MOE. In this figure, the gold region represents the estimate of the MOE for some set of fictional model predictions and field trial observations. The point estimate, perhaps the vector mean value of several similar trials, would be found approximately at the center of this region,

and the overall size of the region represents the uncertainty associated with the point estimate of the MOE.

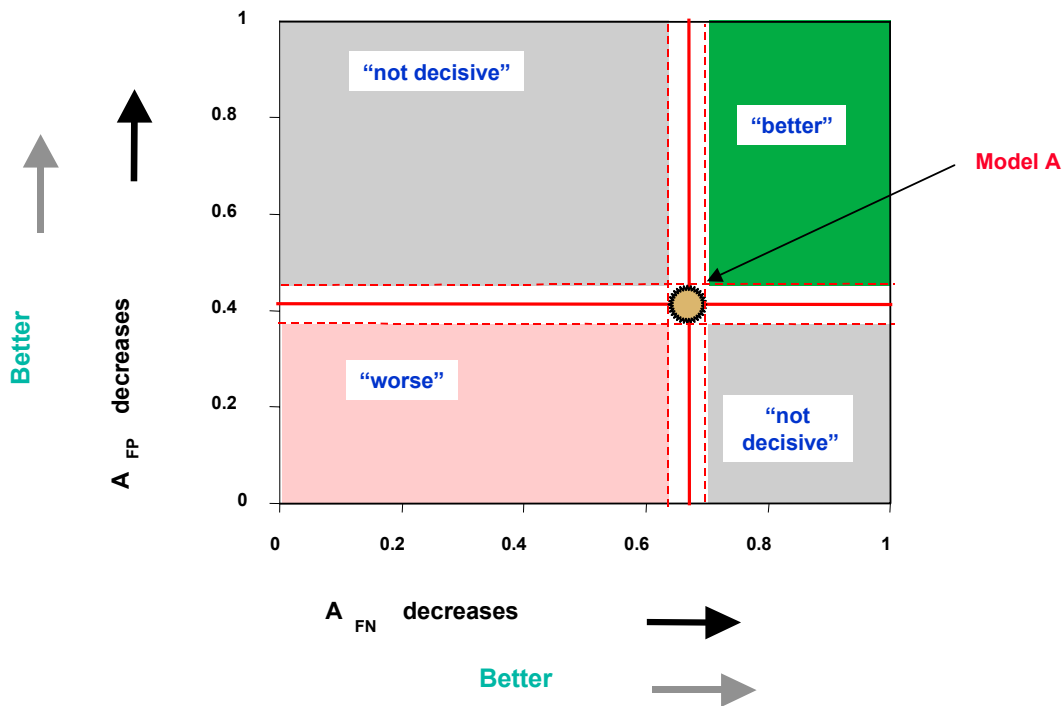


Figure 1-3. Interpretation of Comparisons: Exclusionary Zones

If a second set of model predictions was compared to “Model A,” several conclusions might be anticipated. The second model’s MOE estimate might be found in the region shaded “pink-orange” (lower left). This would imply that Model A performs significantly better; both its false positive and false negative fractions are lower. Alternatively, the second model might lead to an estimate in the green region (upper right) – an indication that Model A is the poorer performer (for this set of field trial observations). Finally, the new model predictions might lead to an MOE value that is located in one of the gray regions. The implication here is that a user would have to make a determination as to the tradeoff between false positive and false negative before deciding which model was most appropriate for his or her specific application.

1. Computation of MOE

Two methods for computing the components of the MOE – A_{OV} , A_{FN} , and A_{FP} – are described in this section. Although Figure 1-1 notionally illustrates physical areas to construct MOE components, the computation of the MOE does not necessarily require

estimated areas and hence, area interpolation. For the analysis reported here, *no area interpolations* were used to compute the MOE values.

a. MOE Based on Concentration (or Dosage)

The components of the MOE – A_{FN} , A_{FP} , and A_{OV} – can be computed directly from the predictions and field trial observations paired in space and time. For the concentration-based MOE, the false positive region is the concentration predicted in a region but not observed. Therefore, for A_{FP} (as shown in Figure 1-4a), one first considers all of the samplers at which the *prediction is of greater value than the observation*. Next, one sums the differences between the predicted and observed concentrations at those samplers. Based on the samplers that contained *observed values that were larger than the predicted values*, one can similarly compute A_{FN} . A_{OV} is calculated by considering all samplers and summing the concentrations associated with the minimum predicted or observed value. Analogous consideration of predicted and observed dosages, results in a summed dosage-based MOE.

b. MOE Based on Concentration or Dosage Threshold

In addition to applying the more general technique described above, one can compute an MOE value based on a prescribed threshold (concentration or dosage). First, one considers the predictions and observations at each of the samplers. If both the prediction and observation are above the threshold, it is considered overlap at that sampler (and the contributions to A_{OV} , A_{FN} , and A_{FP} from this sampler location are 1, 0, 0, respectively). If the prediction is below the threshold and the observation is above, a false negative is assessed at that sampler (and the contributions to A_{OV} , A_{FN} , and A_{FP} from this sampler location are 0, 1, 0, respectively). Similarly, a false positive is assessed when the prediction is above the threshold and the observation is not (and the contributions to A_{OV} , A_{FN} , and A_{FP} from this sampler location are 0, 0, 1, respectively). For the case of a specific sampler at which both the prediction and the observation are below the threshold, the values are assessed as 0,0,0 for the computation of the threshold-based MOE (consistent with the conceptual view illustrated in Figure 1-1). Figure 1-4b illustrates this procedure for a 3-hour average concentration threshold of 0.1 ng m^{-3} . MOE values based on concentration thresholds of 0.01, 0.10, and 0.50 ng m^{-3} were examined in this study. In physical space (given interpolation of observations and predictions), this procedure approximately corresponds to assessing the MOE using a specified contour level (e.g., as illustrated conceptually in Figure 1-1).

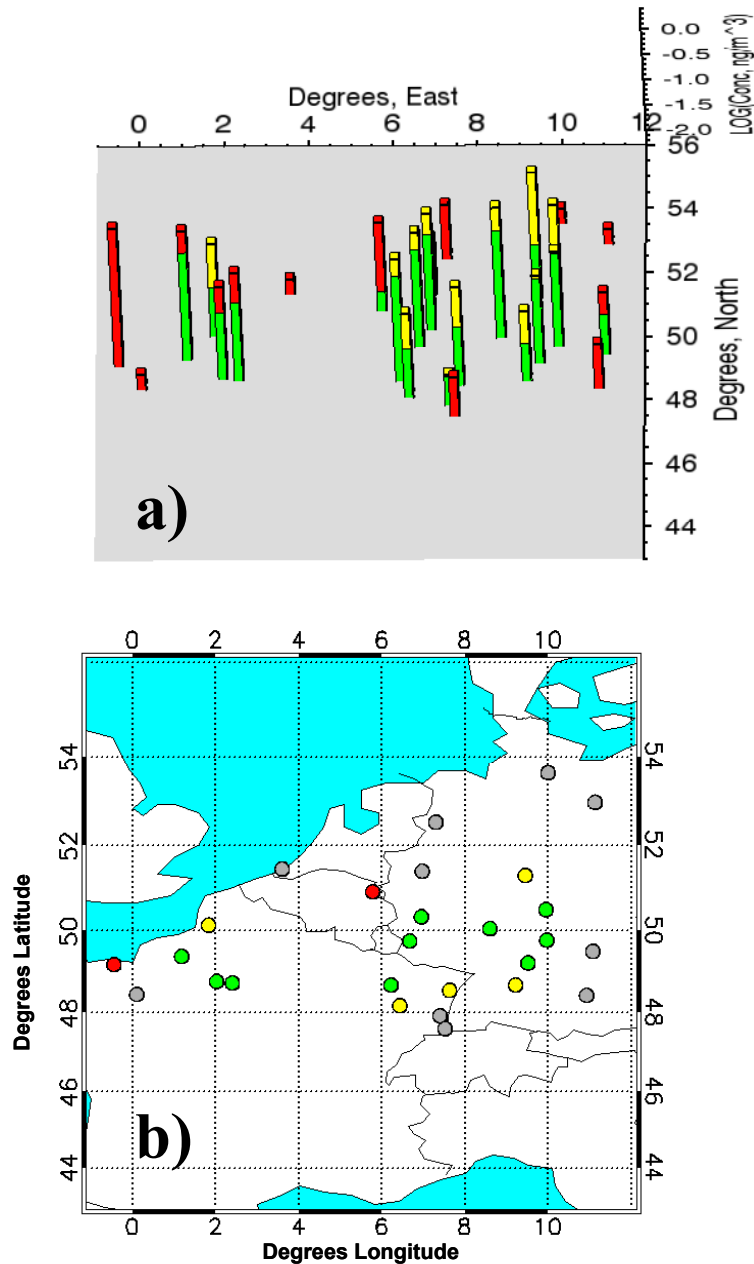


Figure 1-4. Illustration of MOE Computations Based for Model 121 (SCIPUFF) Predictions of the 3-Hour Average Concentrations for the Time Period Between 21 and 24 Hours After the ETEX Release. a) Computation of MOE Based on Summed Concentrations: Green Bars Indicate Overlap, Red Bars Indicate Under-Prediction (“False Negative”), and Yellow Bars Indicate Over-Prediction (“False Positive”). Note that a Logarithmic Scale is Shown. b) Computation of MOE Based on a Threshold of 0.1 ng m^{-3} : Green Circles Indicate Locations Where Both the Observation and the Prediction Were Above 0.1 ng m^{-3} , Red Circles Indicate Locations With an Observation Above 0.1 ng m^{-3} and a Prediction Below 0.1 ng m^{-3} , yellow circles correspond to sampling locations with predictions above 0.1 ng m^{-3} and observations below 0.1 ng m^{-3} and Finally, Gray Circles Indicate That Both the Observation and Prediction Were Below 0.1 ng m^{-3} . Thus, for the portion of Europe Shown in This Example, $A_{OV} = 10$ sampler locations, $A_{FN} = 2$ sampler locations, and $A_{FP} = 5$ sampler locations.

c. Sensitivity of MOE Values to Specific Sampler Location

Previous examinations of transport and dispersion model predictive performance have included estimates of the uncertainty associated with the computed MOE values [Ref. 1-1]. Typically, experiments have included several independent releases (for example, 51 during the *Prairie Grass* experiment and 18 during the *Urban 2000* experiment) that have allowed for the estimation of confidence regions associated with MOE point estimates. In addition, hypothesis test procedures using relatively robust non-parametric techniques have been developed [Ref. 1-9]. For the case of *ETEX*, only one release was examined. We did not attempt to estimate uncertainty bounds given this single release situation in which concentrations are expected to be spatially and temporally correlated. Rather, we assess the variance in computed MOE values by examining time dependence and by considering the influence of any single sampler location.

In order to examine the sensitivity of MOE values to the observation/prediction comparisons of individual locations, we computed MOE values by removing the set of comparisons associated with each location, one at a time. 168 sampling locations across Europe were considered in this analysis. Therefore, for each model we computed 168 “data withheld” (one location at a time) MOE values. The second part of Chapter 2 reports the results of these sensitivity studies.

2. Scoring Functions for the MOE

In this section we develop several notional scoring functions for the MOE space. Essentially, these scoring functions can be thought of as corresponding to the requirements of different possible model users. Such scoring functions can thus aid us in assessing if a model’s MOE value, for a given set of field observations, is “good enough.” In developing the MOE scoring functions, this section also describes and illustrates the mathematical relationships between the figure of merit in space, fractional bias, and a measure of scatter between observations and predictions.

a. Objective Scoring Function: Value Closest to (1,1)

Figure 1-3 suggests that if a given model has a smaller false positive *and* a smaller false negative fraction than some other model, then it is always to be preferred. One can also imagine the situation where a given model’s false positive is decreased but at the expense of an increased false negative, or vice versa. An objective scoring function associated with the MOE space would simply be to consider the values closest to

(1,1) as the best. This scoring approach considers false negative and false positive fractions as *equally* undesirable. For such an objective scoring function (OSF) we define the “distance” to (1,1) – d_{OSF} as

$$d_{OSF} = \sqrt{\left(1 - \frac{A_{OV}}{A_{OB}}\right)^2 + \left(1 - \frac{A_{OV}}{A_{PR}}\right)^2} = \sqrt{\left(\frac{A_{FN}}{A_{OB}}\right)^2 + \left(\frac{A_{FP}}{A_{PR}}\right)^2}. \quad (1-2)$$

Then, for different MOE values, OSF favors the smallest value of d_{OSF} .

b. Risk-Weighted Figure of Merit in Space (RWFMS)

FMS is defined as the ratio of the intersection of the observed and predicted areas to the union of the observed and predicted areas (Eq. 1-3), at a fixed time and above a defined threshold concentration. Reference 1-14 defines FMS as a percentage, and therefore corresponds to Eq. 1-3 multiplied by 100.

$$FMS = \frac{A_{PR} \cap A_{OB}}{A_{PR} \cup A_{OB}}. \quad (1-3)$$

In terms of the MOE nomenclature of false positive, false negative, and overlap regions, the FMS can be rewritten as shown in Eq. 1-4.

$$FMS = \frac{A_{OV}}{A_{OV} + A_{FN} + A_{FP}} \quad (1-4)$$

Importantly, we note that the right-hand side of Eq. 1-4 is actually a more general definition of FMS in that it is not restricted to physical areas, e.g., summing concentrations at all samplers. Now, for the 2D MOE:

$$x = \frac{A_{OV}}{A_{OV} + A_{FN}}, \quad y = \frac{A_{OV}}{A_{OV} + A_{FP}} \quad (1-5)$$

where, for notational convenience, $MOE_x = x$ and $MOE_y = y$. Therefore,

$$A_{FN} = \left(\frac{1-x}{x}\right) A_{OV} \quad \text{and} \quad A_{FP} = \left(\frac{1-y}{y}\right) A_{OV}. \quad (1-6)$$

These definitions of A_{FN} and A_{FP} are then substituted into Eq. 1-4 and following algebraic rearrangement one obtains,

$$FMS = \frac{xy}{xy + y(1-x) + x(1-y)} = \frac{xy}{x + y - xy}. \quad (1-7)$$

Some users of hazardous material transport and dispersion models might consider false positives and false negatives quite differently. For many applications, false positives would be much more acceptable to the user than false negatives (which could result in decisions that directly lead to death or injury). Equation 1-8 is an example of a user scoring function that takes the above risk tolerance into consideration. Basically, this equation describes a modified FMS that includes coefficients, C_{FN} and C_{FP} , to weight the false negative and false positive regions, respectively. We refer to this notional user scoring function as the Risk-Weighted FMS (RWFMS).

$$RWFMS = \frac{A_{OV}}{A_{OV} + C_{FN}A_{FN} + C_{FP}A_{FP}} \quad (1-8)$$

where $C_{FN}, C_{FP} > 0$.

It may be true that, for some applications (e.g., technical model validation), the weightings for false negatives and false positives are considered irrelevant or set equal ($C_{FN} = C_{FP}$). As developed here, the implicit coefficient associated with A_{OV} is 1.0. The precise RWFMS values will depend on the values chosen for C_{FN} and C_{FP} and not just their ratio.

Similar algebraic relationships can be applied to RWFMS to yield the following relationship:

$$RWFMS = \frac{xy}{xy + C_{FN}y(1-x) + C_{FP}x(1-y)} \quad (1-9)$$

Figure 1-5a shows contours of RWFMS (i.e., isolines) in the 2D MOE space for $C_{FN} = C_{FP} = 1$. Similarly, Figure 1-5b illustrates the case where A_{FN} is weighted by a factor of 10 relative to A_{FP} and 5 relative to A_{OV} – i.e., $C_{FN} = 5$ and $C_{FP} = 0.5$.

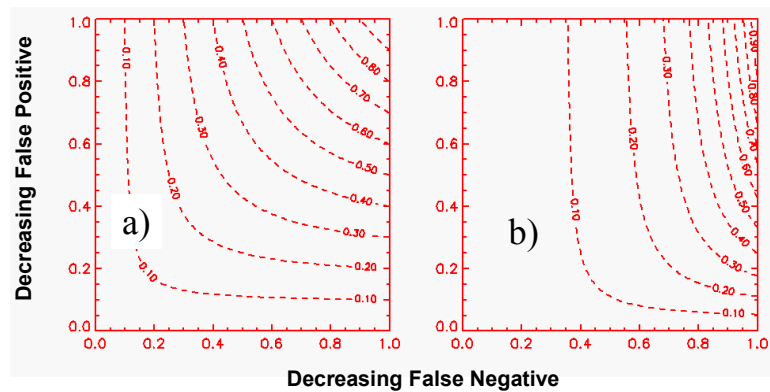


Figure 1-5. Relationship Between RWFMS and 2D MOE, Isolines of RWFMS in the MOE Space: a) $C_{FN} = 1$, $C_{FP} = 1$, b) $C_{FN} = 5$, $C_{FP} = 0.5$

The isolines described in Figure 1-5 can be used as the basis for scoring the performance of a model and for coloring the MOE space according to the RWFMS score. Figure 1-6 provides an example user coloring of the MOE space based on RWFMS for several values of C_{FN} and C_{FP} . At an RWFMS of 0.0, this coloring scheme incorporates pure red. As the user-defined RWFMS increases from 0.0 to 0.50, the intensity of green increases linearly. For instance, at an RWFMS value of 0.5, there are equal intensities of red and green (hence, yellow). Similarly, for RWFMS values between 0.50 and 1.0, the red intensity is reduced linearly with increasing RWFMS value. At an RWFMS value of 1.0, the coloring used is pure green.

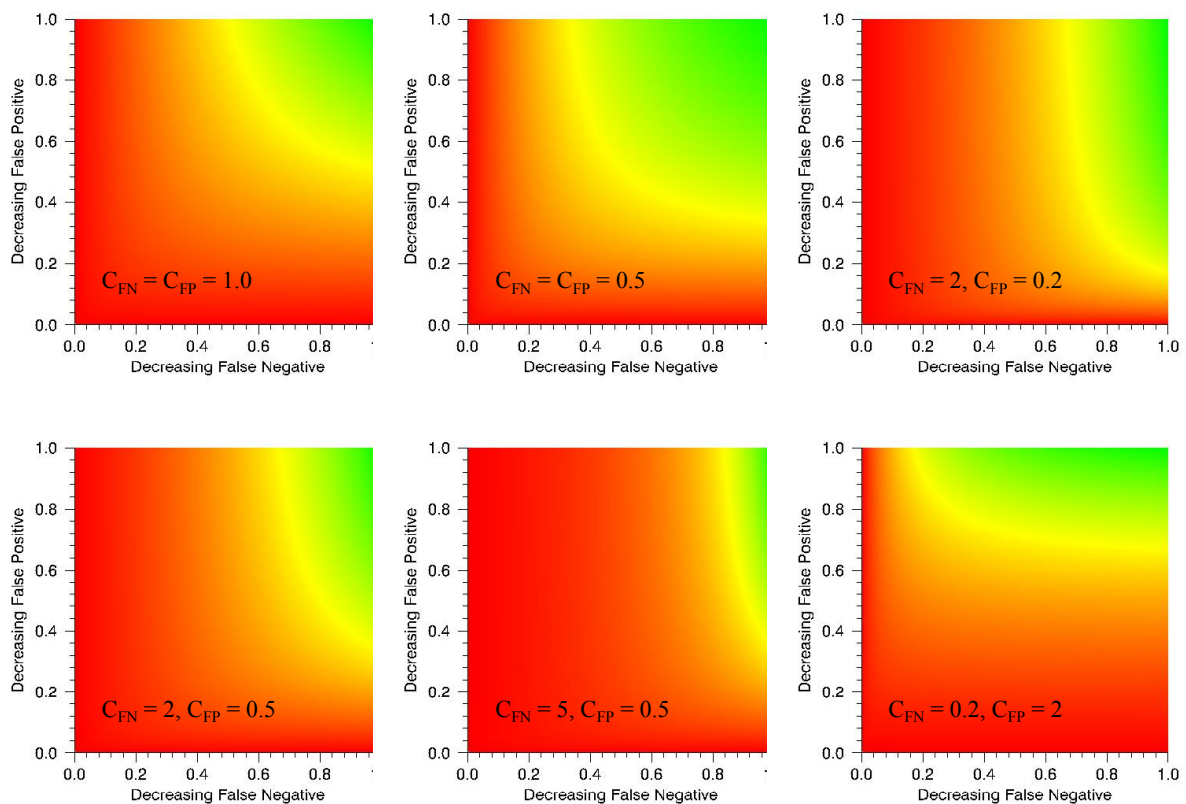


Figure 1-6. User Coloring of MOE Space: RWFMS

c. Fractional Bias Figure of Merit (FBFOM)

A hazardous material transport and prediction model might be applied to problems for which the actual location of the hazard or direction of the plume is of no particular importance. For example, such a model might be used to study potential future

outcomes of an accidental or intentional release. In these cases, the actual weather (e.g., wind speed and direction) of the far future associated with the planning cannot be known with any certainty. For these applications it is desirable to have a scoring function that simply compares the sizes of the predicted and observed areas. In essence, model users in these cases would want a model that minimizes the overall model bias.

Fractional bias (FB), defined below [Ref. 1-15], has been used to evaluate transport and dispersion models under such circumstances;

$$FB = \frac{\overline{C_p} - \overline{C_o}}{0.5(\overline{C_o} + \overline{C_p})}, \quad (1-10)$$

where C = observation/prediction of interest (e.g., dosage), C_p corresponds to model prediction, C_o corresponds to observation, and \overline{C} denotes the average. To begin to explore the relationship between FB and the MOE, we recall that the *summed concentration* 2D MOE is defined as

$$(x, y) = \left(\frac{A_{OV}}{A_{OB}}, \frac{A_{OV}}{A_{PR}} \right). \quad (1-11)$$

Next, consider the ratios

$$\frac{x}{y} = \frac{A_{OV}}{A_{OB}} / \frac{A_{OV}}{A_{PR}} = \frac{A_{PR}}{A_{OB}}. \quad (1-12)$$

We then consider points in the 2D MOE space that lie on the diagonal line, that is, the line $y = x$. Then,

$$1 = \frac{x}{y} = \frac{A_{PR}}{A_{OB}} \Rightarrow A_{PR} = A_{OB}. \quad (1-13)$$

Therefore, this diagonal in the 2D MOE space consists of the points that incorporate “equal size” predictions and observations – no bias, $FB = 0$. Let us assume a hypothetical requirement that A_{PR} and A_{OB} must be within a factor of s of each other, with $s > 1$. Mathematically, this is stated by requiring

$$\frac{1}{s} \leq \frac{A_{PR}}{A_{OB}} \leq s. \quad (1-14)$$

Figure 1-7 plots isolines of this FB figure-of-merit (FBFOM), in MOE space, for various values of parameter s . A coloring scheme (red to green, as discussed previously) for the 2D MOE space using FBFOM can be formulated [Ref. 1-8] with the results shown in Figure 1-8.

One can also relate the fractional bias to the components of the 2D MOE, as follows. From Eq. 1-9 for FB note that:

$$FB = \frac{(\overline{C_p} - \overline{C_o})}{0.5(\overline{C_o} + \overline{C_p})} = \frac{\frac{1}{n} \left(\sum_{i=1}^n C_p^{(i)} - \sum_{i=1}^n C_o^{(i)} \right)}{\frac{1}{2n} \left(\sum_{i=1}^n C_o^{(i)} + \sum_{i=1}^n C_p^{(i)} \right)} = \frac{A_{FP} - A_{FN}}{0.5 \times (2A_{OV} + A_{FN} + A_{FP})}, \quad (1-15)$$

where n = number of data points used in the comparisons and $C_o^{(i)}$ refers to the i^{th} observed concentration, and similarly, $C_p^{(i)}$ refers to the i^{th} predicted concentration. Substituting for A_{FP} and A_{FN} from Eq. 1-6 into Eq. 1-14 leads to

$$FB = \frac{A_{OV} \left(\frac{1-y}{y} \right) - A_{OV} \left(\frac{1-x}{x} \right)}{0.5 \times \left(2A_{OV} + A_{OV} \left(\frac{1-x}{x} \right) + A_{OV} \left(\frac{1-y}{y} \right) \right)} = \frac{\left(\frac{1-y}{y} \right) - \left(\frac{1-x}{x} \right)}{0.5 \times \left(2 + \left(\frac{1-x}{x} \right) + \left(\frac{1-y}{y} \right) \right)} \quad (1-16)$$

and after algebraic simplification

$$FB = \frac{\left(\frac{x - xy - y + xy}{xy} \right)}{0.5 \times \left(\frac{2xy + y - xy + x - xy}{xy} \right)} = \frac{2(x - y)}{x + y}. \quad (1-17)$$

Further rearrangement of Eq. 1-17 yields,

$$y = \frac{2 - FB}{FB + 2} x, \quad (1-18)$$

which shows that isolines of constant FB in the 2D MOE space are straight rays through the origin (Figure 1-7) with slope m

$$m = \frac{y}{x} = \frac{2 - FB}{FB + 2}. \quad (1-19)$$

Within the context of the FB figure of merit, for $FB \geq 0$, $m = 1/s$ from Eq. (1-14) and for $FB < 0$, $m = s$.

The relationships described above for FBFOM and the summed concentration-based MOE are in fact, more general. That is, there is a version of “FB” that is related to the threshold-based MOE in the same mathematical manner as described above. One simply replaces the observed concentrations in Eq. 1-10 with “0,” if the observation is

below the specified threshold, or “1” otherwise. Similarly, for the predictions, “1s” and “0s” replace the predicted concentrations.

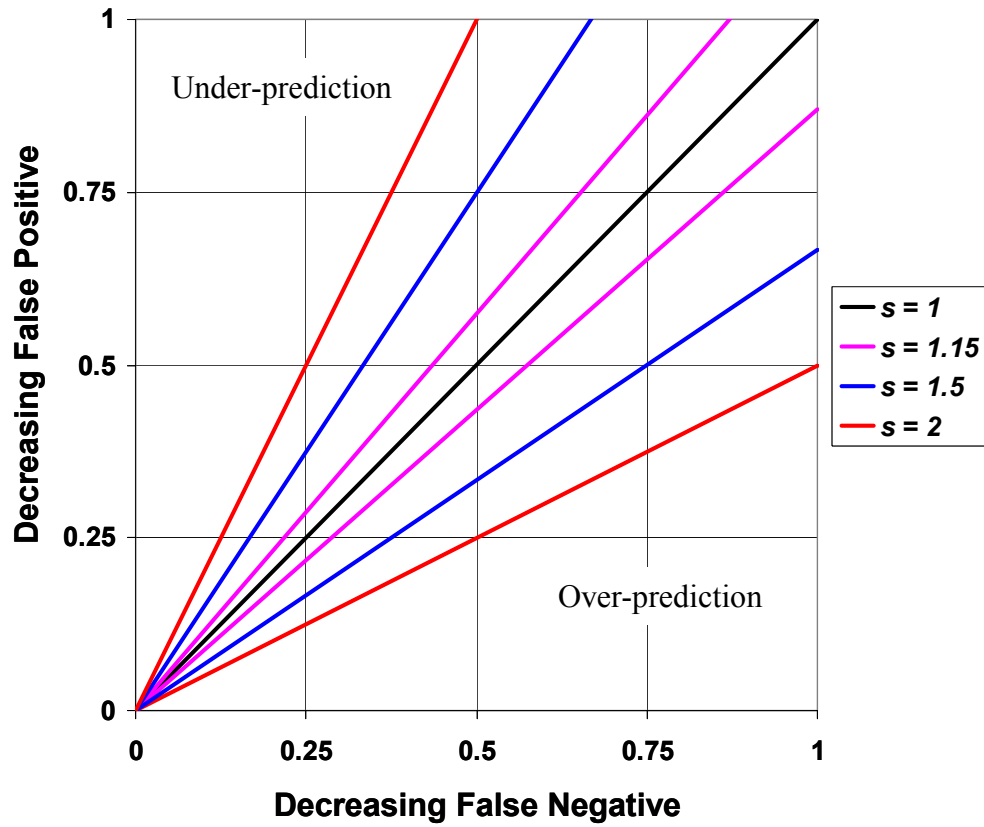


Figure 1-7. FB Isolines for Some Values of the Parameter s

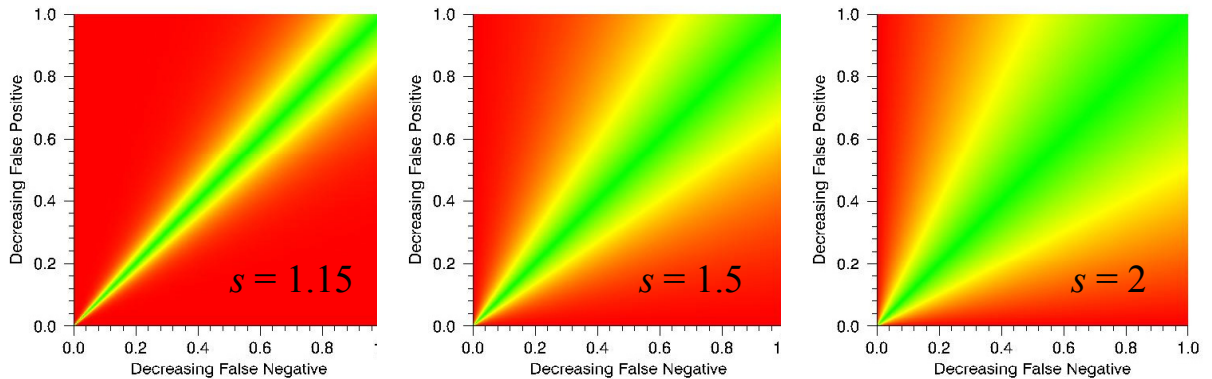


Figure 1-8. Relationship Between FB and MOE: Examples of FBFOM User-Coloring for $s = 1.15, 1.5$ and 2

d. Normalized Absolute Difference (NAD)

Quite often, measures such as mean square error or normalized mean square error are used to characterize the differences between observed and predicted quantities – the scatter if you will. Similar to the way in which bias between a prediction and observation can be portrayed in the MOE space, as discussed above, it is desirable to have a measure of scatter that can be likewise portrayed. For this purpose we define a specialized version of a measure of scatter – normalized absolute difference (NAD) – between observations and predictions:

$$NAD = \frac{\sum_{i=1}^n |C_p^{(i)} - C_o^{(i)}|}{\sum_{i=1}^n (C_o^{(i)} + C_p^{(i)})} \quad (1-20)$$

As with FB, we can express NAD in terms of the false negative, false positive and overlap and, after substitution and algebraic simplification using Eq. 1-6, NAD is related the summed concentration-based MOE components as follows:

$$NAD = \frac{A_{FN} + A_{FP}}{2A_{OV} + A_{FN} + A_{FP}} = \frac{A_{OV}\left(\frac{1-x}{x}\right) + A_{OV}\left(\frac{1-y}{y}\right)}{2A_{OV} + A_{OV}\left(\frac{1-x}{x}\right) + A_{OV}\left(\frac{1-y}{y}\right)} \quad (1-21)$$

$$NAD = \frac{\left(\frac{1-x}{x}\right) + \left(\frac{1-y}{y}\right)}{2A_{OV} + \left(\frac{1-x}{x}\right) + \left(\frac{1-y}{y}\right)} = \frac{\frac{y-xy+x-xy}{xy}}{\frac{2xy+y-xy+x-xy}{xy}} = \frac{x+y-2xy}{x+y} \quad (1-22)$$

Isolines of NAD in the 2D MOE space are shown in Figure 9. Also, NAD is related to FMS (Eq. 1-6) as follows:

$$NAD = \frac{1 - FMS}{1 + FMS}. \quad (1-23)$$

The strictly monotonic relationship between NAD and FMS described by Eq. 1-23 implies that scoring model predictive performance based on NAD or RWFMS (1,1) – the nominal FMS scoring function – will necessarily lead to identical rank orderings. This coincidence with respect to the more natural (operational) FMS scoring function and the more precise NAD scatter scoring function perhaps makes them (NAD or RWFMS (1,1)) even more impressive and valuable.

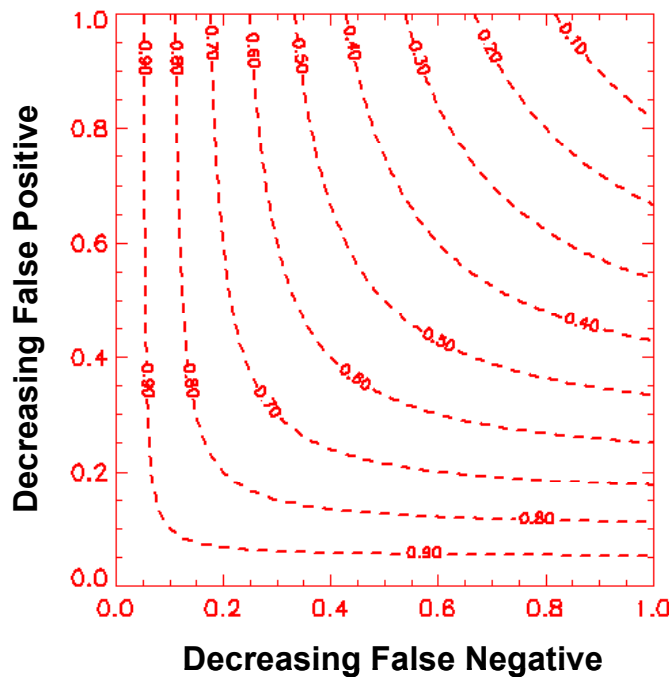


Figure 1-9. Relationship Between NAD and MOE: Isolines of NAD in the 2D MOE Space

As was the case with FBFOM, the mathematical relationship between NAD and the summed concentration-based MOE can be generalized to the threshold-based MOE.

C. BRIEF DESCRIPTION OF THE *EUROPEAN TRACER EXPERIMENT (ETEX)*

The first *ETEX* release, a 12-hour release of the tracer gas perfluoro-methyl-cyclohexane (PMCH), began at 16:00 UTC¹ on 23 October 1994 and ended at 3:50 UTC on 24 October 1994. The release location was 35 km west of Rennes (Monterfil, 20°00'20"W, 48°03'30"N) in Brittany, France. PMCH, an inert, environmentally safe compound, was released 8 m above ground level at a rate of 7.95 g s⁻¹.

Samplers were located at 168 locations across 17 European countries. These samplers were located at synoptic stations of the various national meteorological services. Air samples were collected every 3 hours for a period of 90 hours after the initial release. Figure 1-10 shows the locations of the samplers across Europe.

Measurements of PMCH were made before, during, and after the release at several stations and average background levels were subtracted from the measured data. Furthermore, these measurements suggested that a level of 0.01 ng m⁻³ should be used as

¹ UTC = Universal Time Coordinated.

the minimum for all statistical comparisons. Figure 1-11 illustrates the overall movement and evolution of the “cloud” over a 90-hour sampling period. In Figure 1-11, time progresses by 6-hour increments moving first across columns and then down to the next row.

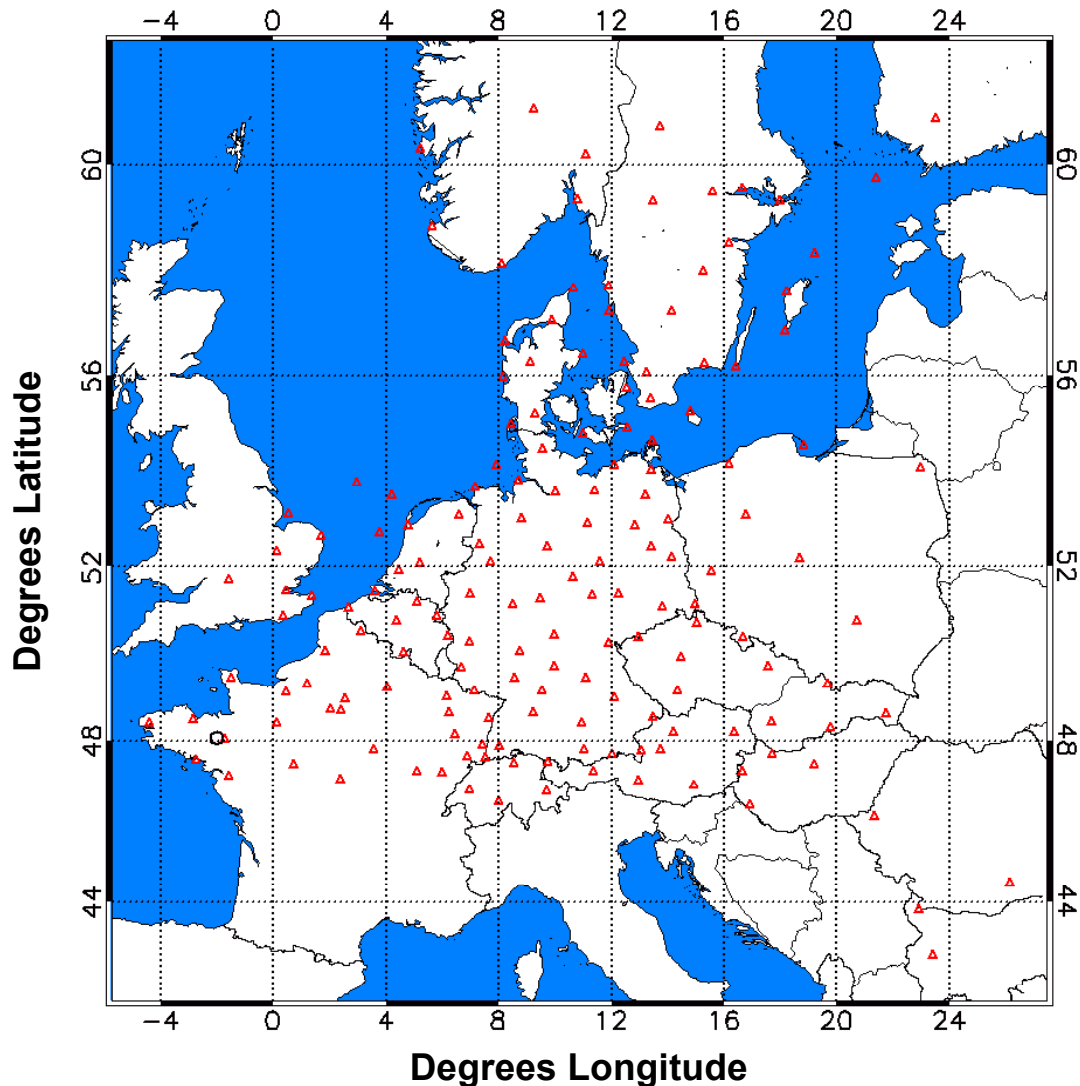


Figure 1-10. ETEX Sampler Locations Across Europe: Red Open Triangles Correspond to Sampler Locations and Black Open Circle Corresponds to the Release Location

The contours of Figure 1-11 are based on an area interpolation procedure. Given values at a discrete (and irregular) set of samplers, the process of interpolation provides intermediate values on some regular grid of points. The resulting regular grid of functional values could be used to obtain contours at specified levels, for instance, of

concentration or dosage. Interpolation procedures can be carried out either in linear or logarithmic space. When interpolating actual plume dosages varying over orders of magnitude, interpolation schemes that use logarithmic space may be considered particularly appropriate.

For the displays of Figure 1-11, we used the Delaunay triangulation procedure. The Delaunay triangulation procedure is useful for the interpolation, analysis, and visual display of irregularly, discretely gridded data. From a set of discrete points (sampler coordinates), a planar triangulation is formed, satisfying the property that the circumscribed circle of any triangle in the triangulation contains no other vertices in its interior.² For any point that is within some triangle (formed via Delaunay triangulation), a linear interpolation routine using values at the vertices of the triangle is used to compute the value at that point. Delaunay triangulation is efficiently implemented in IDL³ and forms a core interpolation routine for display of irregularly gridded data.

We used the above procedure in two ways. First, we used the above procedure directly as described. Next, we first transformed the data (observations and predictions) logarithmically and then followed the above procedure. Both routines were applied with a resolution of 1001×1001 grid points. The displays reported in Figure 1-11 are based on the logarithmic transformation of the data followed by Delaunay triangulation and linear interpolation as described above.

We also briefly examined a few other more complex data fitting routines (e.g., Kriging, Natural Neighbor, Nearest Neighbor, Modified Shepard's, Polynomial Regression, Inverse Distance) [Ref. 1-17]. The resulting "plumes" associated with at least some of these techniques seemed overly sensitive to the adopted parameters associated with the routine and as such, were not used for these qualitative displays. Rather, the adopted Delaunay triangulation procedure followed by linear interpolation, while simple and yielding some perhaps less visually pleasing sharp edges, appeared to be robust and necessarily maintains the actual observed values at the sampler locations (this would not be not true for many fitting procedures).

² Delaunay triangulation is the dual structure of the Voronoi diagram [Ref. 1-16].

³ IDL = Interactive Data Language [Ref. 1-17]. Within IDL, the area interpolation procedure is accomplished by calls to the TRIANGULATE procedure to obtain Delaunay triangulation of the sampler locations followed by the TRIGRID procedure that performs linear interpolation of sampler values to a regular grid.

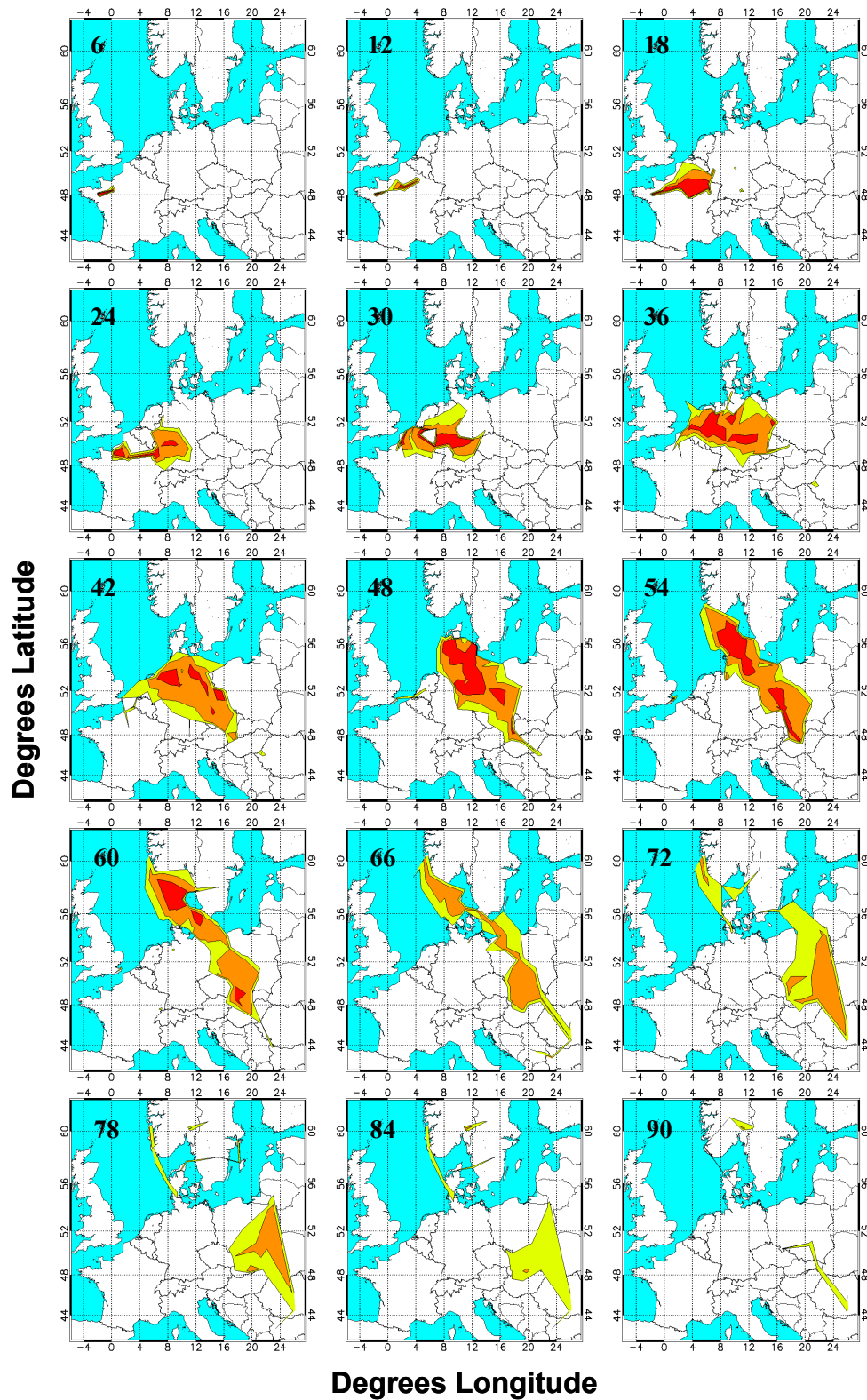


Figure 1-11. Observed PMCH Concentrations Across Europe. Plots Display Contours from 6 Hours After the Release for the Upper Left Plot to 90 Hours After the Release for the Lower Right Plot in Increments of 6 Hours. Contours are 0.01, 0.1, and 0.5 ng m⁻³. Bold numbers on individual plots correspond to the last hour of the given 6-hour period.

Two years after the *ETEX* releases, a modeling exercise known as ATEMS II was conducted.⁴ *ETEX*-ATEMS II predictions associated with 46 model configurations were provided to IDA by the Joint Research Centre (Ispra, Italy), European Commission.⁵ Table 1-1 provides some details associated with these models.⁶ The series of model predictions denoted with a number between 101 and 135, the “100 series,” used European Centre for Medium Range Weather Forecasts (ECMWF) analyzed meteorological data as input. The “200 series” (201-214) used weather inputs selected by the modeler and not the ECMWF-related data. Comparisons between 100 series predictions should tend to identify differences related to variations in dispersion modeling. Comparisons between 200 series predictions or between 100 and 200 series predictions likely will emphasize differences associated with the input wind field. In Table 1-1, model 121, DTRA’s SCIPUFF, and model 127, LLNL’s ARAC, are highlighted in red bold.

Table 1-1. ATEMS II Participants For Which IDA Obtained Predictions

Model	Acronym	Participant	Nationality
101	IMP	Institute of Meteorology and Physics, University of Wien	Austria
102	BMRC	Bureau of Meteorology Research Centre	Australia
103	NIMH-BG	National Institute of Meteorology and Hydrology	Bulgaria
104	NIMH-BG	National Institute of Meteorology and Hydrology	Bulgaria
105	CMC	Canadian Meteorology Centre	Canada
106	DWD	German Weather Service	Germany
107	DWD	German Weather Service	Germany
108	NERI	Nat. Environment Research Inst./Risoe Nat. Lab./Univ. of Cologne	Germany/Denmark
109	NERI	Nat. Environment Research Inst./Risoe Nat. Lab./Univ. of Cologne	Germany/Denmark
110	DMI	Danish Meteorological Institute	Denmark
111	IPSN	French Institute for Nuclear Protection and Safety	France
112	EDF	French Electricity	France
113	ANPA	National Agency for Environment	Italy
114	CNR	National Research Council	Italy
115	JAERI	Japan Atomic Research Institute	Japan

⁴ ATEMS = Atmospheric Transport Model Evaluation Study.

⁵ These predictions were downloaded from the *ETEX* public access web sites: <http://rem.jrc.cec.eu.int/atmes2/> and <http://rem.jrc.cec.eu.int/etex/>.

⁶ An additional three sets of predictions associated with the Royal Dutch Meteorological Institute were not available to us but were part of the original *ETEX* (ATEMS II) study [Ref. 1-12]. This table is extracted from Ref. 1-12.

Table 1-1. ATEMS II Participants For Which IDA Obtained Predictions (continued)

Model	Acronym	Participant	Nationality
116	MRI	Meteorological Research Institute	Japan
117	NIMH-R	National Institute of Meteorology and Hydrology	Romania
118	FOA	Defense Research Establishment	Sweden
119	MetOff	Meteorological Office	United Kingdom
120	NOAA	National Oceanic and Atmospheric Administration	United States
121	ARAP (SCIPUFF)	ARAP Group of Titan Research and Technology	United States
122	KMI	Royal Institute of Meteorology of Belgium	Belgium
123	Meteo	Meteo France	France
127	LLNL (ARAC)	Lawrence Livermore National Laboratories	United States
128	SMHI	Swedish Meteorological and Hydrological Institute	Sweden
129	SAIC	Science Applications International Corporation	United States
130	IMS	Swiss Meteorological Institute	Switzerland
131	DNMI	Norwegian Meteorological Institute	Norway
132	SRS	Westinghouse Savannah River Laboratory	United States
133	JMA	Japan Meteorological Agency	Japan
134	JMA	Japan Meteorological Agency	Japan
135	MSC-E	Meteorological Synthesizing Centre - East	Russia
201	BMRC	Bureau of Meteorology Research Centre	Australia
202	CMC	Canadian Meteorological Centre	Canada
203	DWD	German Weather Service	Germany
204	NERI	Nat. Environment Research Inst./Riseo Nat. Lab./Univ. of Cologne	Germany/Denmark
205	DMI	Danish Metrological Institute	Denmark
206	Meteo	Meteo France	France
207	MRI	Meteorological Research Institute	Japan
208	SMHI	Swedish Meteorological and Hydrological Institute	Sweden
209	MetOff	Meteorological Office	United Kingdom
210	MetOff	Meteorological Office	United Kingdom
211	NOAA	National Oceanic and Atmospheric Administration	United States
212	NIMH-R	National Institute of Meteorology and Hydrology	Romania
213	DNMI	Norwegian Meteorological Institute	Norway
214	MSC-E	Meteorological Synthesizing Centre - East	Russia

D. OUTLINE OF THIS STUDY

Chapter 2 provides the results of this study and includes comparisons of MOE values for the 46 model predictions of the first *ETEX* release that were made available to us. Appendix A lists acronyms and Appendix B provides an extract of the task order associated with this effort.

REFERENCES

- 1-1. Warner, S., Platt, N., and Heagy, J. F., 2004: "User-Oriented Two-Dimensional Measure of Effectiveness for the Evaluation of Transport and Dispersion Models," in press, *J. Appl. Meteor.*
- 1-2. Warner S., Platt, N., and Heagy, 2001: "User-Oriented Measures of Effectiveness for the Evaluation of Transport and Dispersion Models," *Proceedings of the Seventh International Conference on Harmonisation Within Atmospheric Dispersion Modelling for Regulatory Purposes*, Belgirate, Italy, 28-21 May 2001, pages 24-29.
- 1-3. Warner S., Platt, N., Heagy, J. F., Bradley, S., Bieberbach, G., Sugiyama, G., Nasstrom, J. S., Foster, K. T., and Larson, D., 2001: *User-Oriented Measures of Effectiveness for the Evaluation of Transport and Dispersion Models*, IDA Paper P-3554, 797 pp, January 2001. (Available electronically [DTIC STINET ada387239] or on CD via an e-mail request to Steve Warner at swarner@ida.org or a mail request to Steve Warner, Institute for Defense Analyses, 4850 Mark Center Drive, Alexandria, Virginia 22311-1882.)
- 1-4. Warner, S., Platt, N., and Heagy, J. F., 2002: *Explorations to Support the Selection of Trials for HPAC/NARAC Model Comparisons: User-Oriented Measure of Effectiveness Values for the Over-Land Along-Wind Dispersion (OLAD) Field Experiments of September 1997*, IDA Memorandum for DTRA, 24 April 2002.
- 1-5. Platt, N., Warner, S., and Heagy, J. F., 2002: *User-Oriented Measure of Effectiveness Values for Model Predictions of the Transport and Dispersion of Pollutants Inside a Building*, IDA Memorandum for DTRA, 19 June 2002 and Platt N., Warner, S., and Heagy, J. F., "Application of Two-Dimensional User-Oriented Measure of Effectiveness to Interior Building Releases," *Sixth Annual George Mason University Transport and Dispersion Modeling Workshop*, Fairfax, VA, 19 pp, July 2002
- 1-6. Warner, S., Platt N., Heagy, J. F., Bradley, S., Bieberbach, G., Sugiyama, G., Nasstrom, J. S., Foster, K. T., and Larson, D., 2001: "Model Intercomparison with User-Oriented Measures of Effectiveness," *Fifth Annual George Mason University Transport and Dispersion Modeling Workshop*, Fairfax, VA, 15 pp, July 2001.

- 1-7. Warner, S., Heagy, J. F., Platt, N., Larson, D., Sugiyama, G., Nasstrom, J. S., Foster, K. T., Bradley, S., and Bieberbach, G., 2001: *Evaluation of Transport and Dispersion Models: A Controlled Comparison of Hazard Prediction and Assessment Capability (HPAC) and National Atmospheric Release Advisory Center (NARAC) Predictions*, IDA Paper P-3555, 251 pp, May 2001. (Available electronically [DTIC STINET ada391555] or on CD via an e-mail request to Steve Warner at swarner@ida.org or a mail request to Steve Warner, Institute for Defense Analyses, 4850 Mark Center Drive, Alexandria, Virginia 22311-1882.)
- 1-8. Warner, S., Platt, N., and Heagy, 2001: *Application of User-Oriented MOE to HPAC Probabilistic Predictions of Prairie Grass Field*, IDA Paper P-3586, 275 pp, May 2001. (Available electronically [DTIC STINET ada391653] or on CD via an e-mail request to Steve Warner at swarner@ida.org or a mail request to Steve Warner, Institute for Defense Analyses, 4850 Mark Center Drive, Alexandria, Virginia 22311-1882.)
- 1-9. Warner, S. and Platt, N., 2003: *Analyses in Support of Initial Validation of Urban HPAC: Comparisons to Urban 2000 Observations*, IDA Document D-2870N, 119 pp, August 2003. (Available via an e-mail request to Steve Warner at swarner@ida.org or a mail request to Steve Warner, Institute for Defense Analyses, 4850 Mark Center Drive, Alexandria, Virginia 22311-1882.)
- 1-10. Heagy, J. F., Platt, N., and Warner, S., 2003: "Analysis and Measures of Effectiveness Values for Predictions of the DF-5," in *Final Report on the DISCRETE FURY Test Program*, for DTRA by Applied Research Associates, Inc. (Report # ARA-LR-3.03-001), Shock Physics Division, P.O. Box 5388, Albuquerque, NM 87185, August 2003.
- 1-11. Graziani, G., Klug, W., and Mosca, S., 1998: *Real-Time Long-Range Dispersion Model Evaluation of the ETEX First Release*, Joint Research Center, European Commission, Office of Official Publications of the European communities, L-2985 (CL-NA-17754-EN-C), Luxembourg, 1998.
- 1-12. Mosca, S., Bianconi, R., Bellasio, R., Graziani, G., and Klug, W., 1998: *ATEMS II – Evaluation of Long-Range Dispersion Models Using Data of the 1st ETEX Release*, Joint Research Center, European Commission, Office of Official Publications of the European communities, L-2985 (CL-NA-17756-EN-C), Luxembourg, 1998.
- 1-13. Graziani, G., Galmarini, S., Grippa, G., and Klug, W., 1998: *Real-Time Long-Range Dispersion Model Evaluation of the ETEX Second Release*, Joint Research Center, European Commission, Office of Official Publications of the European communities, L-2985 (CL-NA-17755-EN-C), Luxembourg, 1998.
- 1-14. Mosca, S., Graziani, G., Klug, W., Bellasio, R., and Bianconi, R., 1998: "A Statistical Methodology for the Evaluation of Long-Range Dispersion Models: An Application to the ETEX Exercise," *Atmos. Environ.*, 32 (24), 4307-4324.

- 1-15. ASTM, 2000: *Standard guide for statistical evaluation of atmospheric dispersion model performance*. American Society for Testing and Materials (ASTM), D 6589-00, 100 Barr Harbor Dr., PO Box C700, West Conshohocken, PA, 19428, 17 pages.
- 1-16. Guibas, L. J., Knuth, D. E., and Sharir, M., 1992: "Randomized Incremental Construction of Delaunay and Voronoi Diagrams," *Algorithmica* 7: 381-413, 1992. Also, see <http://www.gris.uni-tuebingen.de/gris/proj/dt/dteng.html>.
- 1-17. IDL 6.0, 2003: *Interactive Data Language (IDL)* developed by Research Systems Inc., www.rsinc.com.

CHAPTER 2
RESULTS AND DISCUSSION

2. RESULTS AND DISCUSSION

This chapter describes the results of our analysis. First, MOE values for the 46 *ETEX*-ATEMS II sets of model predictions are presented and compared. In addition, models are ranked using the scoring functions described in Chapter 1. Next, an analysis of the sensitivity of these results – that is, the estimated MOE values – to the influence of any *single* sampler location is described. Finally, this chapter provides an analysis of temporal fluctuations in relative model performance over the 90-hour sampling period that, again, relies on comparisons of computed MOE values.

A. MOE VALUES FOR PREDICTIONS OF *ETEX*

This section provides comparisons of threshold-based (Figure 1-4b) and summed concentration-based (Figure 1-4a) MOE values.

1. Threshold-Based MOE Values: 3-Hour Average Concentration

Figure 2-1 presents the MOE values associated with predictions of 3-hour average concentrations¹ and based on a threshold of 0.01 ng m^{-3} . The MOE values of Figure 2-1 provide information on model performance with respect to predicting the locations of 3-hour average concentrations above 0.01 ng m^{-3} . The numbers in Figure 2-1 correspond to the model number (Table 1-1) with the blue labels referring to the 100 series (e.g., the blue “12” implies model 112) and the red labels referring to the 200 series (e.g., the red “8” implies model 208).

An ellipse has been placed in Figure 2-1 to highlight the result that most of the 46 models led to MOE values in a relatively similar location in the MOE space. Only seven of the model predictions lie outside of this (arbitrary) ellipse (117, 129, 130, 132, 206, 212, and 214). For the 39 model predictions that led to MOE values within the ellipse, it can be seen that they straddle the “45-degree” diagonal (the dashed light purple line). Recall, that an MOE value on this diagonal implies equal sizes of the observed and predicted region – although not necessarily collocation. The variation in MOE

¹ The sample collection time was three hours and thus represents the highest time resolution associated with these data.

performance for the different models within the ellipse appears roughly perpendicular to the diagonal line. The implication of this variation is simply that some models led to over-predictions (those below the diagonal) and some led to under-predictions (those above the diagonal). A few models, like “127” and “204,” resulted in MOE values very near the diagonal, implying little bias in the prediction of the number of locations that exceed the threshold (neither an over- or under-prediction on average).

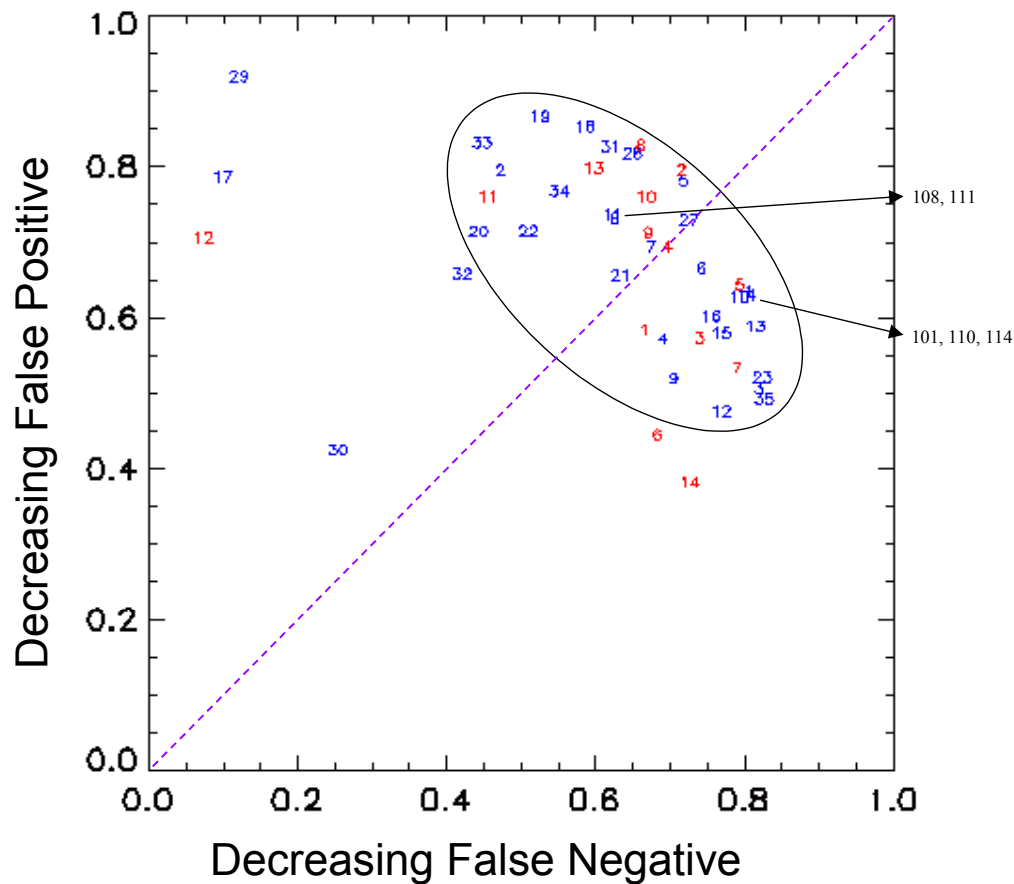


Figure 2-1. 3-Hour Average Concentration Threshold-Based MOE (0.01 ng m^{-3}) Values for 46 ATEMS II Participants. Blue Numbered Labels Refer to Series 100 Models (e.g., “19” implies model 119) and Red Numbered Labels Refer to Series 200 Models (e.g., “13” implies model 213).

Table 1-1 provides rankings of the 46 model predictions shown in Figure 2-1. These rankings are based on 0.01 ng m^{-3} threshold-based MOE values and three scoring functions – the Objective Scoring Function (OSF, “distance” to (1,1)), the Risk-Weighted Figure of Merit in Space (RWFMS) with false negative and false positive weighting coefficients, C_{FN} and C_{FP} , respectively, set to 1, and the RWFMS with the conservative user setting of $C_{\text{FN}} = 5.0$ and $C_{\text{FP}} = 0.5$. See Chapter 1 for scoring function details.

Several features of Table 2-1 can be discussed. First, the rankings associated with OSF and RWFMS (1,1) are quite similar with the first difference in rankings between the two scoring functions occurring at rank 13 and with differences beyond that being relatively minor. Model 121, the SCIPUFF² predictions (that used ECMWF-based weather inputs), were ranked 24 of 46 by both the OSF and the RWFMS (1,1).

The top two models as ranked by OSF and RWFMS (1,1) – 202 and 105 – correspond to CMC (a Canadian model with two different sets of meteorological inputs, see Table 1-1) and the models ranked third and fifth – 208 and 128 – correspond to SMHI (a Swedish model with two different sets of meteorological inputs). The fourth ranked model, 127, was associated with LLNL (Atmospheric Release Advisory Center – ARAC).

The rankings change considerably when the conservative RWFMS (5,0.5) function is used. In this case, model predictions with the smallest false negative fractions, even at the expense of higher false positive fractions, are favored. The C_{FN} and C_{FP} coefficients of 5 and 0.5, respectively, imply that false negative fraction is weighted as 10 times (5/0.5) more important than the false positive fraction. The top ranked model in this case is 113. In Figure 2-1, the 113 MOE value can be seen within the ellipse but toward the right – smaller false negative fraction.

As mentioned in Chapter 1, no measure of uncertainty associated with the MOE point estimates were directly computed (because only a single spatially correlated release was considered). However, we note the following. During our previous examination of Urban HPAC model predictions of *Urban 2000* [Ref. 2-1], we found that, because we had eighteen independent releases to examine, we could generate reasonable confidence regions associated with the MOE point estimates. In addition, non-parametric hypothesis test procedures were used to identify statistically significant differences between model predictive performances. In one case, we found that although we could rank the model performance (as in Table 2-1), the predictions of several of the 20 Urban HPAC model configurations could not be statistically distinguished based our hypothesis testing. For example, of 20 models, we could statistically distinguish the performance of thirteen from the top performer [Ref. 2-2]. That is, the predictive performance of the other six Urban HPAC model configurations could not be statistically distinguished from the top-performing model (at least by a single metric).

² SCIPUFF = Second-Order Closure Integrated Puff. SCIPUFF is the main transport and dispersion model within DTRA's current Hazard Prediction and Assessment Capability (HPAC).

Table 2-1. Relative Model Rankings Based on 0.01 ng m⁻³ Threshold MOE Values for Three Scoring Functions: OSF, RWFMS (1,1) and RWFMS (5,0.5)

Rank	Model	OSF	Model	RWFMS (1,1)	Model	RWFMS (5,0.5)
1	202	0.358	202	0.597	113	0.402
2	105	0.361	105	0.594	101	0.401
3	208	0.388	208	0.574	114	0.396
4	127	0.389	127	0.568	123	0.394
5	128	0.397	128	0.565	135	0.388
6	210	0.413	210	0.548	103	0.384
7	131	0.420	131	0.546	110	0.384
8	101	0.420	101	0.545	205	0.381
9	205	0.420	205	0.544	207	0.355
10	114	0.424	114	0.541	115	0.348
11	106	0.427	106	0.537	116	0.338
12	110	0.431	110	0.535	106	0.334
13	204	0.439	118	0.530	112	0.327
14	118	0.441	204	0.526	127	0.325
15	209	0.445	209	0.521	105	0.322
16	107	0.451	107	0.517	202	0.316
17	213	0.453	213	0.516	203	0.314
18	113	0.457	113	0.514	204	0.289
19	111	0.463	111	0.507	109	0.281
20	108	0.464	108	0.506	104	0.277
21	116	0.472	116	0.500	107	0.276
22	115	0.485	115	0.489	210	0.274
23	119	0.494	119	0.485	214	0.270
24	121	0.507	121	0.472	209	0.269
25	134	0.508	134	0.471	208	0.268
26	203	0.508	203	0.470	128	0.263
27	123	0.516	123	0.464	201	0.254
28	207	0.519	207	0.461	206	0.249
29	103	0.532	104	0.452	121	0.240
30	104	0.533	103	0.450	131	0.240
31	201	0.542	201	0.445	108	0.239
32	135	0.543	135	0.441	111	0.238
33	102	0.568	109	0.423	213	0.222
34	109	0.569	122	0.422	118	0.217
35	122	0.570	102	0.419	134	0.192
36	112	0.578	112	0.412	119	0.179
37	133	0.579	133	0.409	122	0.167
38	211	0.597	211	0.397	102	0.149
39	120	0.629	120	0.374	211	0.140
40	206	0.648	206	0.363	133	0.138
41	132	0.675	132	0.344	120	0.133
42	214	0.681	214	0.329	132	0.123
43	129	0.883	130	0.188	130	0.061
44	117	0.927	129	0.120	129	0.027
45	130	0.945	117	0.097	117	0.022
46	212	0.974	212	0.072	212	0.016

The point of the previous paragraph is simply a warning. Although Table 2-1 (and the upcoming Tables 2-2 through 2-5) shows a ranking of model performance for *ETEX* predictions, one should be aware that some of the differences between models are likely not statistically significant.

Figure 2-2 presents MOE values for the 46 models when a threshold of 0.1 ng m^{-3} is applied. 40 models lead to MOE values within the (arbitrary) circular region highlighted in Figure 2-2 and again, the predictions of seven models (the same seven as in Figure 2-1) led to MOE values outside the region. 27 of 46 MOE values lie below the diagonal, indicating over-prediction with respect to the number of locations with 3-hour average concentrations above 0.1 ng m^{-3} .

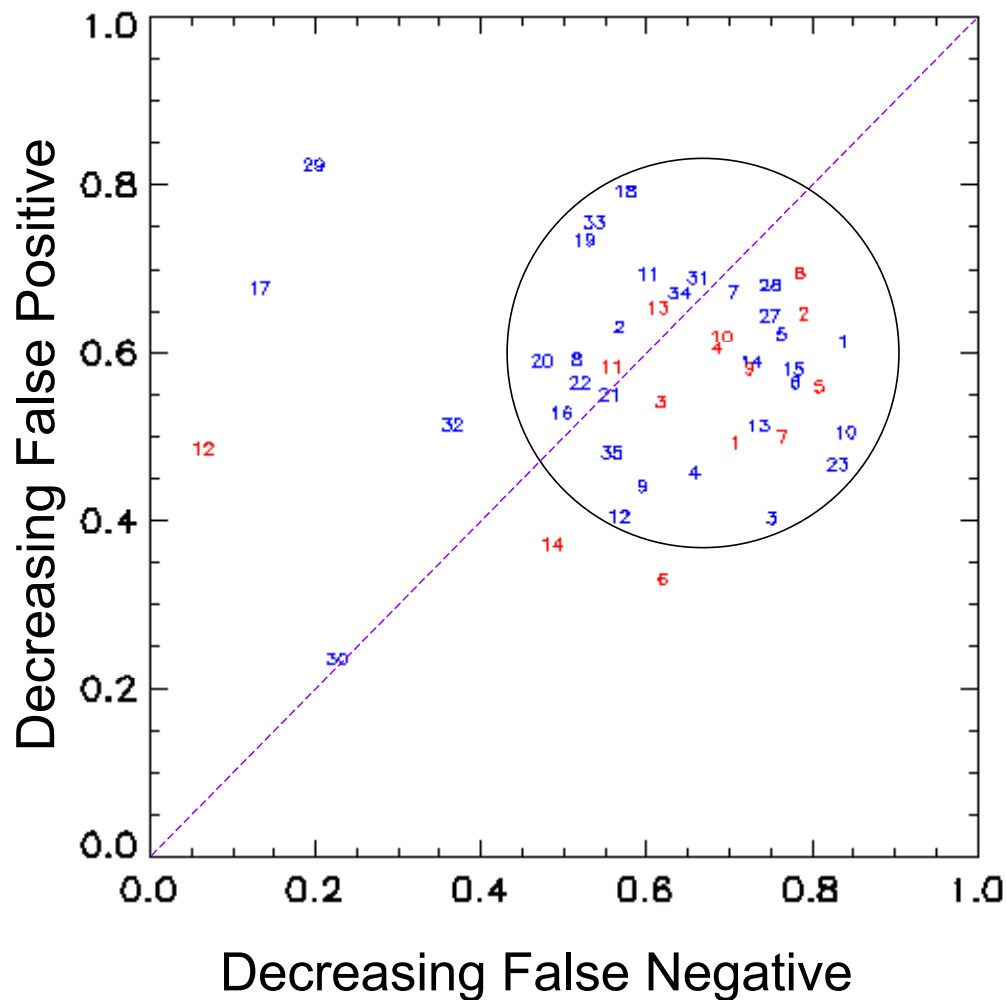


Figure 2-2. 3-Hour Average Concentration Threshold-Based MOE (0.1 ng m^{-3}) Values for 46 ATEMS II Participants. Associated Table Presents Relative Model Rankings for Three Functions: OSF, RWFMS (1,1) and RWFMS (5,0.5). Blue Numbered Labels Refer to Series 100 Models (e.g., “19” implies model 119) and Red Numbered Labels Refer to Series 200 Models (e.g., “13” implies model 213).

Table 2-2 lists the rankings for the 46 models based on MOE values computed with a 0.1 ng m^{-3} threshold and for the OSF, RWFMS (1,1) and RWFMS (5,0.5) scoring functions. As was true in Table 2-1, the rankings based on OSF and RWFMS (1,1) shown in Table 2-2 are quite similar. For example, model 121 (SCIPUFF) ranked 30 and model 127 (LLNL-ARAC) ranked 5 by both scoring functions (OSF and RWFMS (1,1)). Model 101, (IMP, an Austrian model) achieves a very high ranking for both RWFMS (1,1) and RWFMS (5,0.5) scoring functions (i.e., 4th and 1st, respectively).

MOE values based on 3-hour concentrations and a threshold of 0.5 ng m^{-3} are shown in Figure 2-3. Overall, it can be seen that the MOE values have degraded substantially, that is, they have moved away from (1,1). This relatively high threshold value-based MOE favors models that predict the locations/times of the higher concentrations (and hence the shorter times after the release and closer downwind distances). The implication is that predicting the locations and times of the higher 3-hour average concentrations is a challenge to all models.

Figure 2-3 indicates that most of the 46 models over-predicted the number of locations (times) with 3-hour average concentrations above 0.5 ng m^{-3} . Table 2-3 provides the associated rankings based on the MOE values of Figure 2-3. Staying with our previous examples, we note that SCIPUFF ("121") ranked 23 and 21 by OSF and RWFMS (1,1), respectively and LLNL-ARAC ("127") ranked first by both scoring functions. Using the RWFMS (5,0.5) scoring function dropped the relative ranking of SCIPUFF to 28 and LLNL-ARAC to 6.

Table 2-2. Relative Model Rankings Based on 0.1 ng m⁻³ Threshold MOE Values for Three Scoring Functions: OSF, RWFMS (1,1) and RWFMS (5,0.5)

Rank	Model	OSF	Model	RWFMS (1,1)	Model	RWFMS (5,0.5)
1	208	0.381	208	0.577	101	0.438
2	128	0.411	128	0.551	110	0.409
3	202	0.419	202	0.545	123	0.385
4	101	0.424	101	0.544	205	0.381
5	127	0.440	127	0.526	208	0.380
6	107	0.446	107	0.521	202	0.377
7	105	0.451	105	0.517	115	0.357
8	131	0.462	131	0.508	106	0.357
9	118	0.476	118	0.497	105	0.350
10	115	0.481	115	0.493	128	0.344
11	205	0.488	205	0.487	127	0.339
12	134	0.492	134	0.484	207	0.323
13	106	0.494	106	0.482	114	0.309
14	210	0.495	210	0.481	113	0.305
15	114	0.499	114	0.478	209	0.301
16	111	0.505	111	0.473	107	0.300
17	209	0.509	209	0.470	103	0.294
18	204	0.513	204	0.467	210	0.282
19	213	0.522	213	0.460	201	0.275
20	110	0.526	110	0.456	204	0.271
21	133	0.526	133	0.455	131	0.265
22	119	0.547	119	0.439	134	0.246
23	113	0.560	113	0.428	104	0.238
24	207	0.562	207	0.426	213	0.227
25	123	0.565	102	0.423	111	0.221
26	102	0.572	123	0.421	203	0.217
27	201	0.594	201	0.403	118	0.207
28	203	0.606	203	0.399	109	0.198
29	211	0.612	211	0.396	102	0.196
30	121	0.637	121	0.379	206	0.193
31	108	0.638	108	0.378	211	0.188
32	104	0.647	122	0.368	121	0.185
33	103	0.652	104	0.365	133	0.183
34	122	0.653	120	0.354	135	0.182
35	120	0.671	103	0.351	112	0.180
36	135	0.687	135	0.345	119	0.176
37	116	0.694	116	0.341	122	0.166
38	109	0.695	109	0.336	108	0.166
39	112	0.741	112	0.306	116	0.154
40	206	0.778	132	0.270	120	0.145
41	132	0.803	206	0.269	214	0.140
42	214	0.817	214	0.263	132	0.099
43	129	0.822	129	0.190	130	0.051
44	117	0.927	130	0.129	129	0.047
45	212	1.071	117	0.125	117	0.030
46	130	1.092	212	0.060	212	0.014

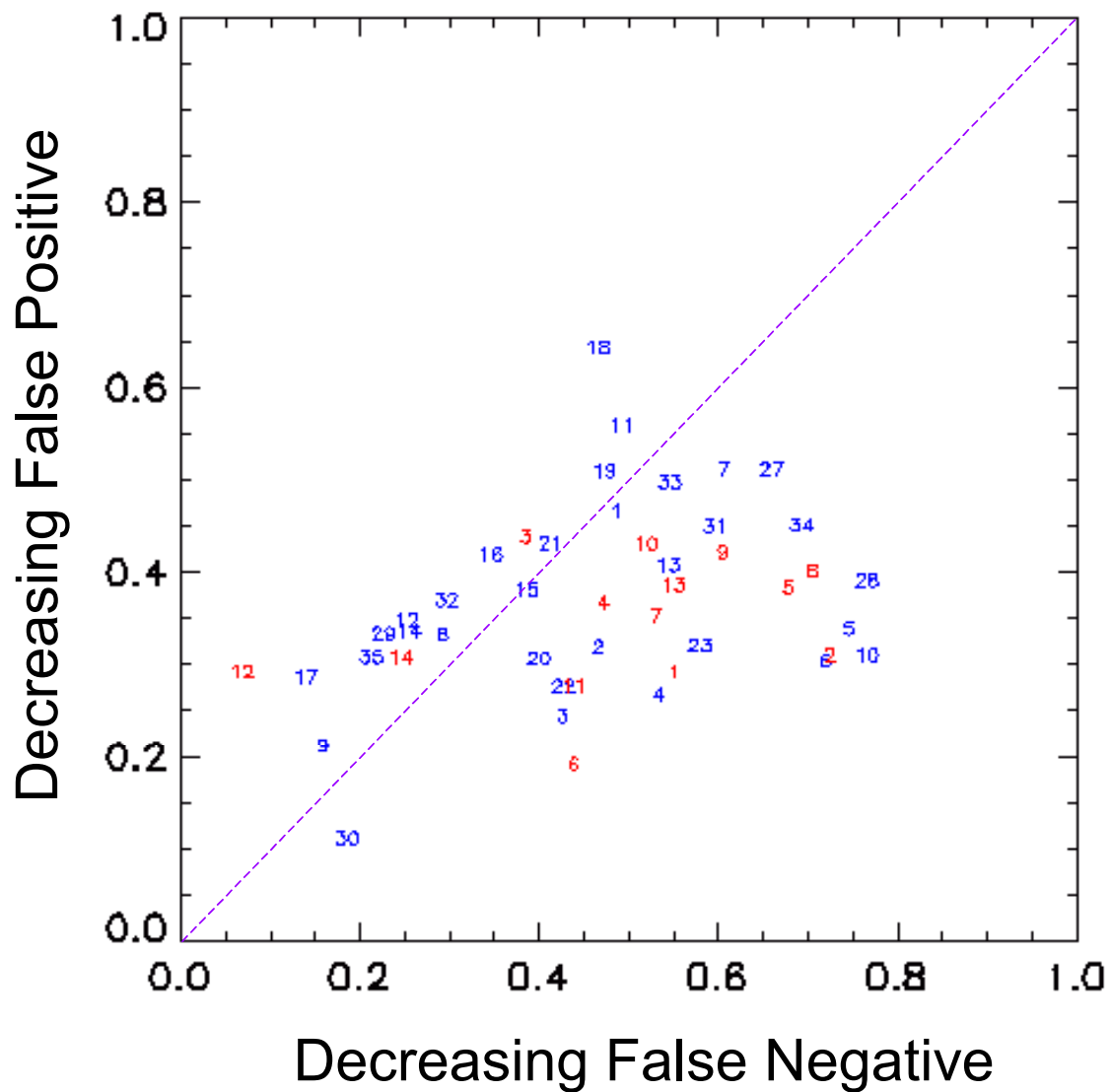


Figure 2-3. 3-Hour Average Concentration Threshold-Based MOE (0.5 ng m^{-3}) Values for 46 ATEMS II Participants. Associated Table Presents Relative Model Rankings for Three Functions: OSF, RWFMS (1,1) and RWFMS (5,0.5). Blue Numbered Labels Refer to Series 100 Models (e.g., “19” implies model 119) and Red Numbered Labels Refer to Series 200 Models (e.g., “13” implies model 213).

Table 2-3. Relative Model Rankings Based on 0.5 ng m⁻³ Threshold MOE Values for Three Scoring Functions: OSF, RWFMS (1,1) and RWFMS (5,0.5)

Rank	Model	OSF	Model	RWFMS (1,1)	Model	RWFMS (5,0.5)
1	127	0.600	127	0.401	128	0.301
2	107	0.632	107	0.381	110	0.273
3	134	0.635	134	0.371	105	0.270
4	118	0.646	118	0.368	134	0.261
5	128	0.658	111	0.352	208	0.256
6	208	0.676	133	0.349	127	0.246
7	111	0.676	128	0.344	202	0.245
8	133	0.682	131	0.341	106	0.242
9	131	0.687	208	0.338	205	0.235
10	205	0.704	209	0.325	107	0.211
11	209	0.709	119	0.323	131	0.200
12	105	0.714	205	0.319	209	0.198
13	119	0.724	101	0.309	133	0.177
14	110	0.735	210	0.305	123	0.175
15	101	0.746	113	0.300	213	0.169
16	210	0.750	105	0.299	113	0.169
17	202	0.750	213	0.290	210	0.159
18	113	0.752	110	0.278	201	0.156
19	106	0.756	202	0.272	207	0.155
20	213	0.766	106	0.267	111	0.153
21	123	0.804	121	0.265	104	0.147
22	207	0.807	207	0.264	101	0.146
23	121	0.822	123	0.256	118	0.143
24	204	0.832	204	0.255	119	0.142
25	203	0.841	203	0.253	204	0.132
26	201	0.846	115	0.235	102	0.128
27	102	0.870	116	0.232	211	0.115
28	104	0.874	201	0.231	121	0.114
29	115	0.876	102	0.230	122	0.111
30	116	0.879	104	0.212	103	0.107
31	211	0.920	120	0.207	120	0.103
32	120	0.922	211	0.201	206	0.103
33	122	0.928	122	0.198	115	0.102
34	132	0.948	132	0.196	203	0.102
35	103	0.954	108	0.183	116	0.090
36	108	0.977	103	0.180	132	0.073
37	206	0.993	112	0.170	108	0.071
38	112	0.995	114	0.168	114	0.060
39	114	1.002	214	0.157	112	0.060
40	129	1.025	129	0.155	214	0.057
41	214	1.028	206	0.148	129	0.052
42	135	1.052	135	0.143	135	0.049
43	117	1.122	117	0.103	130	0.037
44	109	1.156	109	0.099	109	0.034
45	212	1.173	130	0.072	117	0.030
46	130	1.210	212	0.059	212	0.014

2. Summed Concentration-Based MOEs

Figure 2-4 presents the MOE values computed based on comparing observed and predicted 3-hour average concentrations (see Figure 1-4a). 39 of the 46 models led to an over-prediction (i.e., their MOE value lies below the diagonal). The predictions of model 121, SCIPUFF, led to the largest false positive fraction. This over-prediction has been discussed in the past [Ref. 2-4] and is, in large part, due to an over-prediction at a single sampler location. This will be described in more detail in the next section.

Table 2-4 provides rankings for the 46 model predictions of 3-hour average concentrations based on five scoring functions. As was done for the threshold-based MOE values, models are ranked based on OSF, RWFMS (1,1) and RWFMS (5,0.5). In addition, fractional bias (FB) and Normalized Absolute Difference (NAD) are used to rank values. The absolute value of FB is actually used for this ranking. There is a one-to-one correspondence (in the mathematical sense) between NAD and RWFMS (1,1) and thus the rankings are identical. RWFMS (1,1) and NAD provide for very similar rankings to OSF, with the first six ranked models being identical.

Table 2-4 reports that the model with the least fractional bias was 203 (DWD, a German model) followed by 107 (which was also DWD, but using a differing set of meteorological input). Figure 2-4 also illustrates this result – 107 and 203 are closest to the diagonal. Model 107 is also ranked highest when using OSF and RWFMS (1,1) (and hence, necessarily, NAD). For these two DWD model predictions, 107 led to improved performance as judged by the MOE in spite of the fact that the two model configurations led to very similar bias performance, suggesting the improvements were related to the meteorological inputs used for 107 (relative to 203).

This section has demonstrated that MOE values can be computed and interpreted for comparisons of predictions and observations associated with *very long-range* (many hundreds of kilometers) transport and dispersion experiments. MOE values based on threshold concentrations and summed concentrations have been presented. The use of scoring functions to rank model performance has also been demonstrated.³ The objective scoring function (OSF) allows one to rank model performance based on how close the model's MOE value is to (1,1). The other scoring functions that have been used can aid assessments of relative model performance for different nominal applications (for

³ The rankings reported in the tables of this section are (when appropriately compared) reasonably consistent with those reported previously in Ref. 2-3.

example, a conservative hazard warning area application might use the RWFMS (5,0.5) scoring).

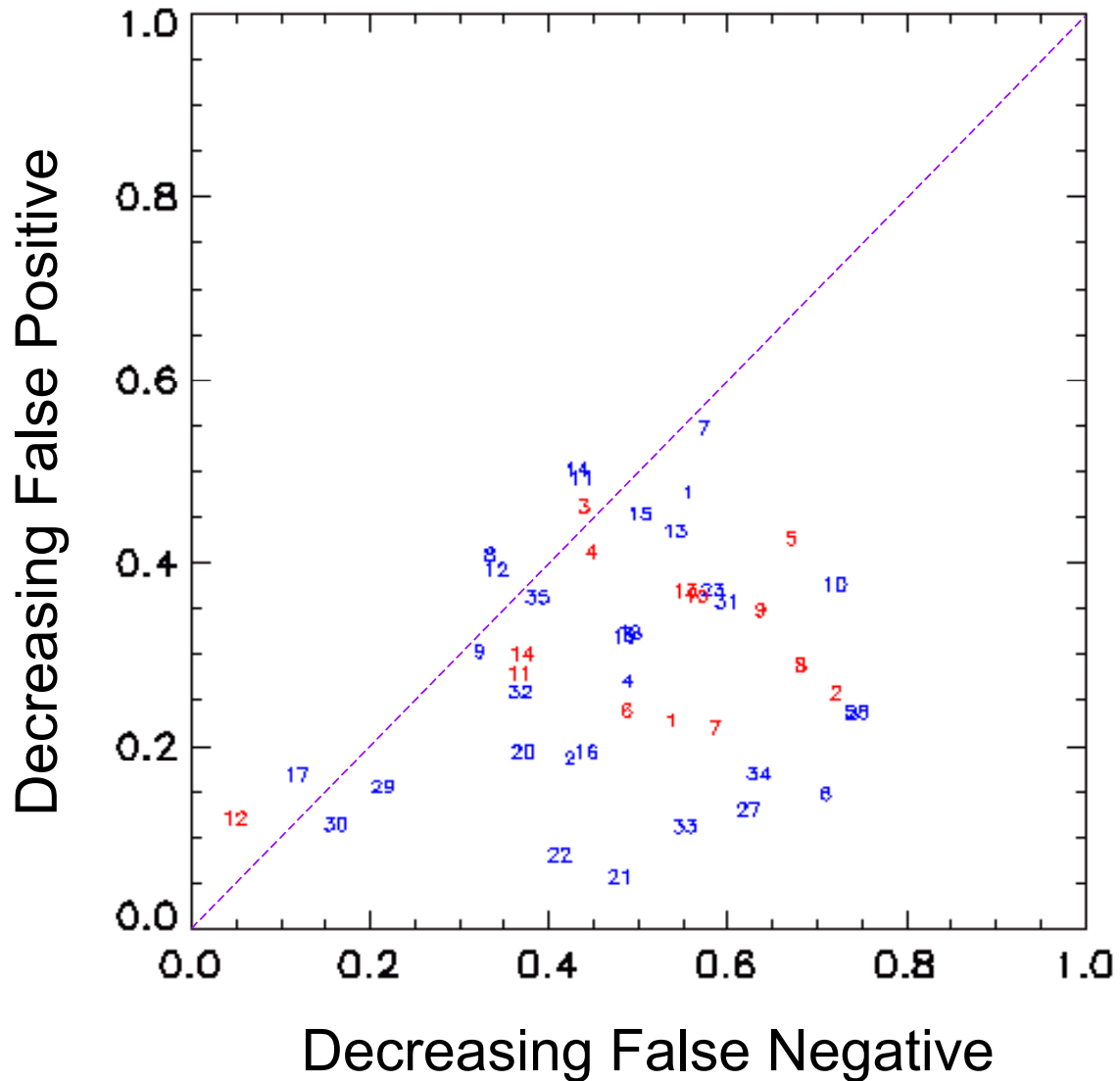


Figure 2-4. MOE Values for 46 ATEMS II Participants and Based on 3-Hour Average Concentrations Comparisons. Blue Numbered Labels Refer to Series 100 Models (e.g., “12” implies model 112) and Red Numbered Labels Refer to Series 200 Models (e.g., “8” implies model 208).

Table 2-4. Relative Model Rankings Based on Summed Concentration MOE Values for Five Scoring Functions: OSF, RWFMS (1,1), NAD, RWFMS (5,0.5), and FB (i.e., absolute value of FB)

Rank	Model	OSF	Rank	Model	RWFMS (1,1)	NAD	Rank	Model	RWFMS (5,0.5)	Rank	Model	ABS(FB)
1	107	0.625	1	107	0.387	0.442	1	110	0.264	1	203	0.046
2	205	0.669	2	205	0.347	0.485	2	205	0.238	2	107	0.061
3	110	0.689	3	101	0.343	0.490	3	128	0.229	3	109	0.083
4	101	0.690	4	110	0.324	0.511	4	105	0.225	4	135	0.085
5	113	0.733	5	113	0.314	0.522	5	202	0.224	5	204	0.086
6	115	0.744	6	115	0.310	0.527	6	208	0.215	6	111	0.105
7	209	0.754	7	114	0.300	0.539	7	209	0.205	7	115	0.117
8	123	0.760	8	111	0.299	0.539	8	107	0.195	8	112	0.122
9	114	0.760	9	123	0.289	0.552	9	131	0.190	9	114	0.134
10	111	0.762	10	209	0.285	0.556	10	123	0.184	10	101	0.169
11	131	0.762	11	203	0.285	0.556	11	101	0.181	11	108	0.184
12	210	0.776	12	131	0.284	0.557	12	210	0.174	12	214	0.226
13	213	0.778	13	213	0.281	0.561	13	113	0.169	13	113	0.235
14	203	0.786	14	210	0.280	0.562	14	213	0.169	14	211	0.299
15	208	0.787	15	204	0.269	0.577	15	106	0.164	15	117	0.302
16	202	0.800	16	208	0.249	0.602	16	207	0.155	16	129	0.365
17	128	0.810	17	118	0.239	0.615	17	134	0.155	17	132	0.374
18	105	0.812	18	103	0.237	0.617	18	115	0.152	18	130	0.406
19	204	0.815	19	119	0.236	0.619	19	201	0.140	19	213	0.414
20	118	0.852	20	202	0.229	0.627	20	118	0.138	20	119	0.430
21	103	0.854	21	135	0.228	0.629	21	103	0.137	21	118	0.431
22	119	0.857	22	108	0.224	0.635	22	119	0.135	22	103	0.436
23	135	0.888	23	112	0.221	0.637	23	127	0.132	23	210	0.452
24	207	0.890	24	128	0.214	0.648	24	104	0.131	24	205	0.453
25	108	0.894	25	105	0.214	0.648	25	111	0.126	25	123	0.464
26	104	0.896	26	104	0.207	0.657	26	206	0.125	26	131	0.524
27	112	0.900	27	214	0.196	0.672	27	204	0.124	27	209	0.595
28	106	0.907	28	206	0.186	0.687	28	114	0.124	28	104	0.596
29	201	0.908	29	211	0.186	0.687	29	203	0.123	29	110	0.643
30	134	0.914	30	207	0.186	0.687	30	133	0.107	30	120	0.659
31	206	0.926	31	201	0.185	0.687	31	116	0.105	31	206	0.700
32	214	0.946	32	109	0.183	0.690	32	135	0.102	32	212	0.784
33	127	0.954	33	132	0.176	0.700	33	102	0.099	33	102	0.809
34	211	0.964	34	116	0.151	0.737	34	214	0.093	34	116	0.812
35	109	0.976	35	134	0.149	0.741	35	211	0.092	35	208	0.824
36	132	0.979	36	102	0.145	0.747	36	132	0.091	36	201	0.824
37	116	0.986	37	120	0.142	0.751	37	112	0.088	37	207	0.923
38	133	1.001	38	106	0.133	0.766	38	120	0.086	38	202	0.960
39	102	1.002	39	127	0.115	0.794	39	108	0.086	39	105	1.045
40	120	1.027	40	133	0.097	0.823	40	109	0.079	40	128	1.055
41	121	1.083	41	129	0.096	0.824	41	122	0.070	41	134	1.186
42	122	1.096	42	117	0.073	0.863	42	121	0.063	42	106	1.340
43	129	1.158	43	130	0.069	0.871	43	129	0.045	43	127	1.340
44	117	1.217	44	122	0.067	0.874	44	130	0.032	44	133	1.361
45	130	1.225	45	121	0.047	0.909	45	117	0.025	45	122	1.389
46	212	1.300	46	212	0.036	0.931	46	212	0.010	46	121	1.623

B. SENSITIVITY OF MOE VALUES TO SINGLE SAMPLER LOCATIONS

A previous analysis of SCIPUFF predictions of the *ETEX* release [Ref. 2-4] suggested that a large model over-prediction at one sampler located near the release point greatly impacted some of the metrics used to assess model performance. The particular sampler with the large over-prediction was “F21” located near Rennes, France. For example, this previous report found that, when including all samplers, a normalized mean square error of 2160 was computed but after removing the single sampler at Rennes, a value of 14.8 was obtained. Similar sensitivity was observed for measures of bias.

We explore this sensitivity in this section by computing MOE values by removing one sampler location at a time. That is, since there are 168 sampler locations, one computes a total of 169 MOE values; one value that included all sampler locations and 168 values that considered only 167 samplers with a single location removed for each. We examine these values to see how variable the MOE estimates are and identify those locations that have a large impact on the MOE value. Clearly, samplers near the release point, where concentrations are expected to be highest, should have the greatest influence on MOE values.

First, we should note that MOE values based on a threshold should be relatively robust to the removal of a single sampler location. This is necessarily true because, recall (Figure 1-5), that the components of the threshold-based MOE value – overlap, false negative, and false positive – are computed by simply counting the number of samplers in which predictions and observations are above the threshold (overlap), observations are above and predictions are below (false negative), and predictions are above and observations are below (false positive). Therefore, the removal of a single location will lead to a change in only a few sampler counts associated with overlap, false negative, and false positive. Figure 2-5 illustrates this robust behavior. The center of the large red diamond in Figure 2-5 corresponds to the nominal (all sampler locations included) 0.1 ng m^{-3} threshold-based MOE estimate for model 121, SCIPUFF. The smaller 168 black “plus signs” (seen as a close black cluster) correspond to the “one location at-a-time removed” MOE values. It can be seen that little variance in the threshold-based MOE values can be associated with the removal of any single sampler location. For example, compare the variance in 0.1 ng m^{-3} threshold-based MOE estimates shown in Figure 2-5 with the variance between different models shown in Figure 2-2. This robust behavior (to the removal of a single sampler location) was true for all 46 sets of model predictions and for the three threshold-values (0.01 , 0.1 , and 0.5 ng m^{-3}) that we examined. The conclusion, then, is that relative model performance as assessed by threshold-based MOE values is not sensitive to the removal of a single sampler location.

Next, we consider the MOE values base on summed concentrations (as in Figure 1-3). In this case, we expect that MOE values are potentially sensitive to sampler locations near the release where the concentrations will be highest. Of the 46 model predictions that we examined, the summed concentration-based MOE values of most were relatively unaffected by the removal of any single sampler. Figure 2-6 shows typical results for six models.

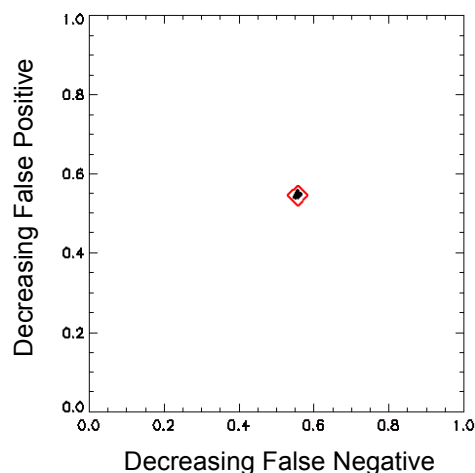


Figure 2-5. MOE Values for SCIPUFF (Model 121) Predictions of *ETEX* and Based on a 0.10 ng m^{-3} Threshold for 3-Hour Average Concentrations Comparisons. Small Black “Plus Signs” (in a Close Cluster) Correspond to the 168 MOE Values That Are Computed After the Removal of One Sampling Location at a Time; The Center of the Large Red Diamond Corresponds to the Nominal (All Sampler Locations Included) MOE Value.

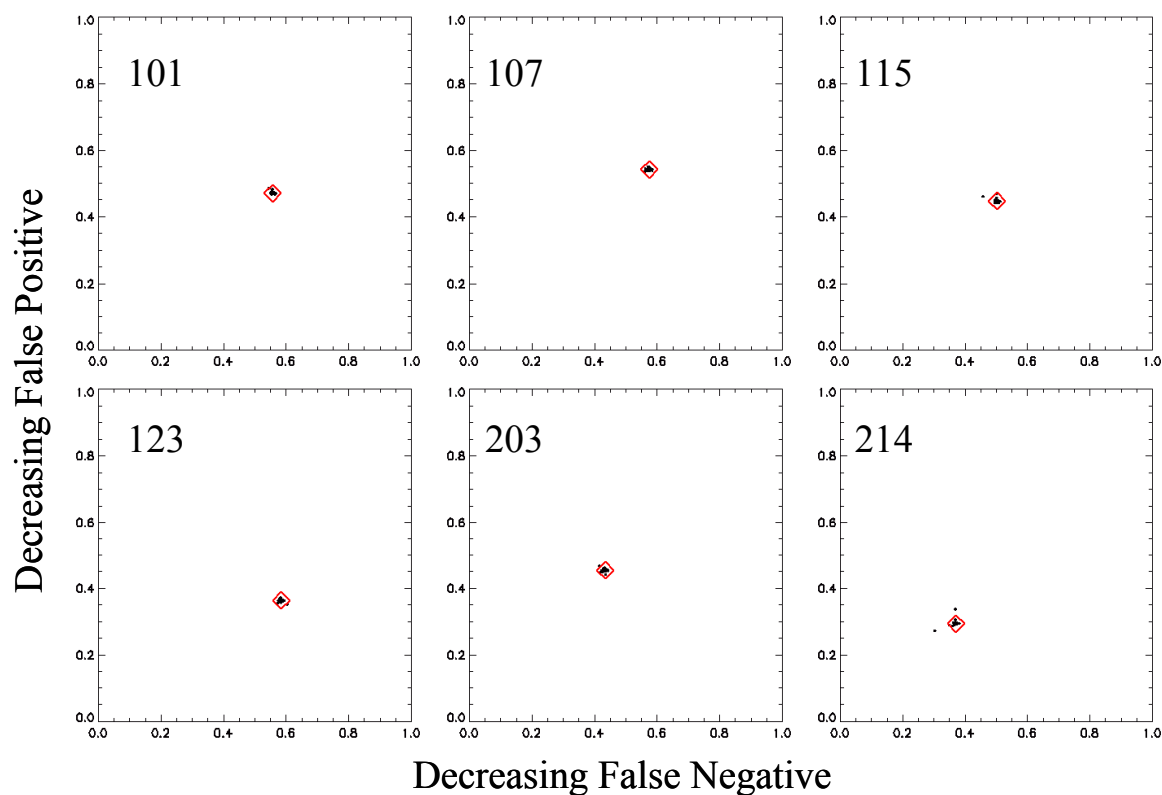


Figure 2-6. Summed Concentration-Based MOE Values for Six Model Predictions of *ETEX* for 3-Hour Average Concentrations Comparisons. Models are as Follows: 101 = IMP, Austria; 107 = DWD, Germany; 115 = JAERI, Japan; 123 = Meteo, France; 203 = DWD Germany; and 214 = MSC-E, Russia. (See Table 1-1 for more details.) Small Black Points (Clustered “Plus Signs”) Correspond to the 168 MOE Values That Are Computed After the Removal of One Sampling Location at a Time; The Nominal (All Sampler Locations Included) MOE Value Lies at the Center of the Large Red Diamond.

The six models that were most sensitive to the removal of a single sampler location were 116 (MRI, Japan), 118 (GOA, Sweden), 121 (ARAP [SCIPUFF], United States), 127 (LLNL [ARAC], United States), 132 (SRS, United States), and 210 (MetOff, United Kingdom). In these cases, the removal of a single sampler location caused a substantial change in the estimated MOE value. Figure 2-7 illustrates this result for these six models. In all six cases, the sampler location that had the substantial influence was “F21,” located in Rennes, France near the release location. Figure 2-8 shows the F21 sampler location and the release location. In all cases, the removal of F21 results in far less over-prediction by these models. That is, these six models greatly over-predicted the concentrations associated with F21. Other models, including most of the 100-series predictions that used the same ECMWF weather inputs, did not have such a large over-prediction at F21.

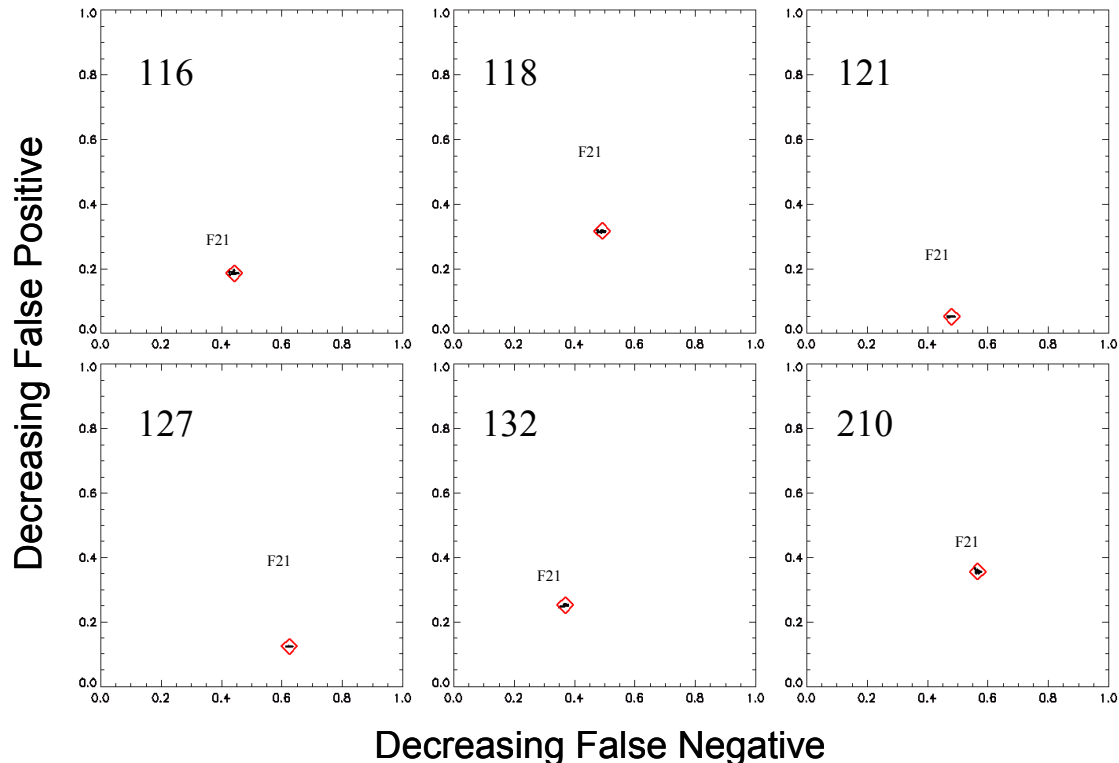


Figure 2-7. Summed Concentration-Based MOE Values for the Six Model Predictions of *ETEX* for 3-Hour Average Concentrations Comparisons That Were Most Influenced by a Single Sampler Location. Models are as Follows: 116 = MRI, Japan; 118 = FOA, Sweden; 121 = ARAP [SCIPUFF], United States; 127 = LLNL [ARAC], United States; 132 = SRS, United States; and 210 = MetOff, United Kingdom. (See Table 1-1 for more details.) The Nominal (All Sampler Locations Included) MOE Value Lies at the Center of the Large Red Diamond, Small Black Points (Clustered “Plus Signs”) Correspond to 167 MOE Values That Are Computed After the Removal of One Sampling Location at a Time, and the Point Labeled “F21” Corresponds to the MOE Value Computed When Location F21 is Removed.

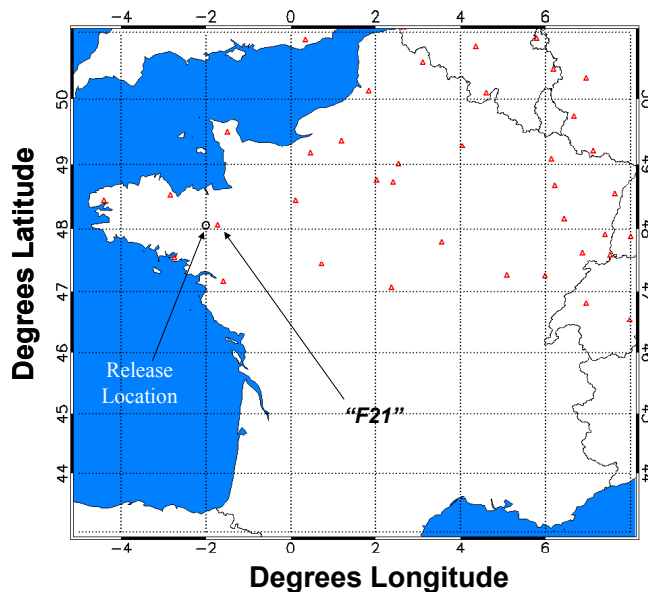


Figure 2-8. Sampler F21 (Rennes) and Release Point (Monterfil) Locations

Reference 2-4 suggested the following explanation for the observed SCIPUFF (Model 121) performance:

“The meteorological data provided for the calculation also lacked a boundary layer description, so the early plume development is strongly influenced by the model choice for the boundary layer. Since SCIPUFF does not routinely accept the ECMWF data as input, and no special effort was made to develop an interface, the boundary layer description was relatively uncertain.”

Table 2-5 presents relative model rankings, as before in Table 2-4, but after the exclusion of sampler location F21 from the computation of all MOE values. Several of the top rankings remain similar to those in Table 2-4. For example, models 101, 107, 110, and 205 remain in the top 5 based on scoring functions OSF, RW-FMS (1,1), and NAD when using all sampler locations (Table 2-4) and when excluding F21 (Table 2-5). Table 2-5 also shows significant changes in the ranks associated with the six models that were identified as having MOE values that were sensitive to the inclusion of F21 – 116, 118, 121, 127, 132, and 210. For SCIPUFF (121), previous rankings of 41, 45, and 46, for the OSF, RW-FMS (1,1) or NAD, and ABS(FB) scoring functions were associated with the inclusion of all sampler locations (Table 2-4). Removing the F21 sampler location results in SCIPUFF rankings becoming 34, 34, and 30 (Table 2-5) for the OSF, RW-FMS(1,1) or NAD, and ABS(FB) scoring functions, respectively. The biggest

relative improvements for any model predictions were associated with ARAC (127). ARAC rankings went from 33, 39, and 43 (all samplers, Table 2-4) to 8, 8, and 22 (with F21 removed, Table 2-5) for the OSF, RW-FMS(1,1) or NAD, and ABS(FB) scoring functions, respectively. Models 116, 118, 132, and 210 show improvements in relative rankings similar to those seen for SCIPUFF (121).

Table 2-5. Relative Model Rankings Based on Summed Concentration MOE Values for Five Scoring Functions After the Removal of Sampler Location F21: OSF, RWFMS (1,1), NAD, RWFMS (5,0.5), and FB

Rank	Model	OSF	Rank	Model	RWFMS (1,1)	NAD	Rank	Model	RWFMS (5,0.5)	Rank	Model	ABS(FB)
1	107	0.641	1	107	0.376	0.453	1	110	0.255	1	115	0.002
2	101	0.689	2	101	0.344	0.488	2	205	0.228	2	203	0.007
3	205	0.689	3	205	0.333	0.501	3	128	0.225	3	112	0.049
4	118	0.707	4	118	0.330	0.504	4	105	0.212	4	107	0.052
5	110	0.709	5	111	0.324	0.511	5	202	0.210	5	111	0.055
6	111	0.722	6	210	0.318	0.518	6	208	0.209	6	108	0.078
7	210	0.729	7	110	0.308	0.529	7	209	0.196	7	214	0.111
8	127	0.733	8	127	0.308	0.529	8	123	0.193	8	101	0.119
9	209	0.736	9	209	0.305	0.533	9	127	0.192	9	135	0.121
10	113	0.753	10	113	0.301	0.537	10	107	0.187	10	119	0.122
11	131	0.755	11	115	0.298	0.541	11	131	0.177	11	132	0.132
12	123	0.760	12	131	0.295	0.544	12	101	0.175	12	210	0.176
13	115	0.765	13	204	0.286	0.555	13	106	0.171	13	109	0.189
14	208	0.772	14	123	0.286	0.556	14	210	0.164	14	204	0.192
15	204	0.781	15	203	0.281	0.561	15	113	0.163	15	114	0.192
16	203	0.794	16	208	0.268	0.578	16	213	0.162	16	118	0.244
17	213	0.801	17	213	0.266	0.580	17	201	0.161	17	113	0.258
18	128	0.806	18	114	0.261	0.586	18	207	0.154	18	117	0.260
19	202	0.818	19	119	0.260	0.587	19	134	0.150	19	116	0.309
20	114	0.825	20	108	0.240	0.613	20	104	0.150	20	131	0.390
21	105	0.828	21	103	0.233	0.622	21	204	0.145	21	211	0.403
22	119	0.828	22	112	0.228	0.628	22	111	0.143	22	127	0.415
23	103	0.855	23	128	0.221	0.638	23	103	0.143	23	129	0.433
24	207	0.861	24	207	0.219	0.640	24	115	0.133	24	213	0.449
25	108	0.867	25	202	0.218	0.642	25	118	0.130	25	209	0.451
26	104	0.870	26	104	0.214	0.647	26	203	0.124	26	205	0.472
27	201	0.879	27	135	0.212	0.650	27	119	0.123	27	130	0.493
28	112	0.889	28	105	0.204	0.661	28	206	0.119	28	103	0.509
29	106	0.893	29	116	0.198	0.669	29	102	0.113	29	123	0.530
30	134	0.917	30	211	0.196	0.673	30	121	0.110	30	121	0.550
31	135	0.918	31	109	0.194	0.675	31	211	0.105	31	110	0.682
32	211	0.937	32	132	0.194	0.675	32	133	0.103	32	207	0.683
33	116	0.938	33	201	0.192	0.678	33	114	0.102	33	212	0.687
34	121	0.943	34	121	0.186	0.686	34	116	0.100	34	104	0.689
35	206	0.947	35	206	0.172	0.706	35	108	0.098	35	208	0.701
36	109	0.951	36	214	0.168	0.713	36	120	0.098	36	120	0.752
37	132	0.953	37	134	0.154	0.732	37	112	0.094	37	206	0.756
38	102	0.976	38	106	0.152	0.737	38	109	0.090	38	102	0.896
39	120	1.002	39	102	0.150	0.740	39	135	0.083	39	201	0.911
40	214	1.007	40	120	0.148	0.742	40	132	0.075	40	202	0.970
41	133	1.013	41	133	0.100	0.818	41	214	0.073	41	128	1.003
42	122	1.119	42	129	0.077	0.857	42	122	0.063	42	105	1.054
43	129	1.203	43	117	0.068	0.873	43	129	0.037	43	134	1.118
44	117	1.232	44	130	0.063	0.882	44	130	0.031	44	106	1.234
45	130	1.237	45	122	0.062	0.883	45	117	0.023	45	133	1.291
46	212	1.296	46	212	0.039	0.926	46	212	0.011	46	122	1.381

C. MOE VALUES AS A FUNCTION OF TIME

The MOE values computed and discussed to this point have considered all locations (i.e., where samplers were “turned on”) and all times (i.e., out to 90 hours after the release). For example, the summed concentration MOE was computed based on comparing predictions and observations for all of the 30 3-hour average concentration periods. In order to examine the performance of a set of model predictions as a function of time after the release, one can consider computing MOE values for each of the individual 3-hour time periods – there are 30 such time periods or “strokes” for this *ETEX* data set.

MOE values for these strokes were calculated and, in addition, MOE values were computed in a cumulative manner (i.e., for the first 3 hours, for the first 6 hours, for the first 9 hours, and so on, all the way to 90 hours which was what was presented in the last section). Next, MOE values for “running time windows” were computed. For this technique, running time windows (RTW) of 12 and 24 hours were examined. For the 12-hour RTW, the first MOE value computed is based on the first four 3-hour strokes and the next value is based on strokes “2” through “5,” the next is based on strokes “3” through “6,” and so on. Similarly, for the 24-hour RTW computations, six 3-hour strokes were used for the computation of each time-dependent MOE value.⁴

Examples of the above time-dependent MOE values are shown in Figure 2-9, illustrated with the results for model 121 (SCIPUFF) and model 127 (ARAC). All of the MOE values shown in Figure 2-9 are based on a threshold of 0.01 ng m^{-3} . The top two plots of Figure 2-9 present the 30 threshold-based MOE values computed for each of the 3-hour strokes. The line connects each of these values in the correct time order (from “START” which corresponds to the first 3-hour period to the last point which corresponds to the last 3-hour period – i.e., the MOE value based on the prediction of the 3-hour average concentration associated with the time period between 88 and 90 hours after the release). The middle two plots show threshold-based MOE values based on 12-hour RTW with MOE values colored blue corresponding to the “independent” values that occur every 12 hours. Similarly, the bottom two plots present threshold-based MOE values based on 24-hour RTW and show the corresponding “independent” values in blue for every eighth point.

⁴ Although not described in any detail here, similar time-dependent MOE values were also created for each model after the removal of one sampler location at a time.

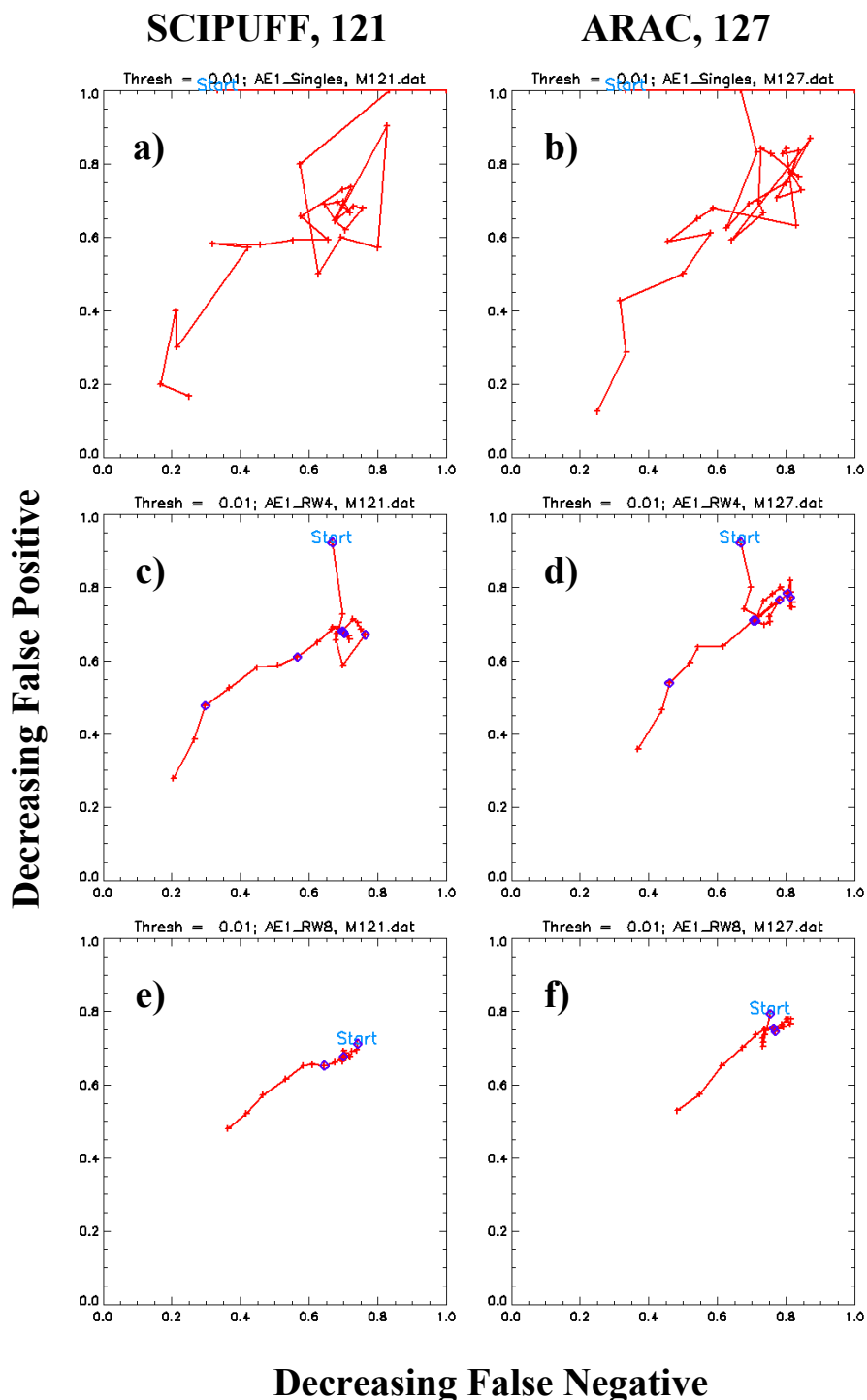


Figure 2-9. MOE Values Based on a Threshold Concentration of 0.01 ng m⁻³ for Model Predictions of *ETEX*. a) Values for SCIPUFF for the 30 Consecutive 3-Hour Periods, b) Values for ARAC for the 30 Consecutive 3-Hour Periods, c) Values for SCIPUFF for 12-Hour RTW, d) Values for ARAC for 12-Hour RTW, e) Values for SCIPUFF for 24-Hour RTW, and f) Values for ARAC for 24-Hour RTW. For the 12-Hour RTW Plots, MOE Values Colored Blue Diamonds Correspond to the “Independent” Values That Occur Every 12 Hours. For the 24-Hour RTW Plots, MOE Values Colored Blue Diamonds Correspond to the “Independent” Values That Occur Every 24 Hours.

In Figure 2-9 it can be seen that the 0.01 ng m^{-3} MOE values degrade with time. That is, both the false negative and false positive fractions get larger as a function of increased time. The MOE values appear to move down, at least roughly, the “45-degree” diagonal as a function of time. This suggests, that for both SCIPUFF and ARAC, the overall size of the prediction might be about right (at least when considering a 0.01 ng m^{-3} threshold) but the actual predicted locations do not correspond to the observed locations and this mismatch gets worse with time. This degradation in time thus appears to be related to model transport (wind direction and speed, as opposed to dispersion) of PMCH. This behavior, degradation in model predictive performance as a function of time (and distance), appears for many of the sets of model predictions.

Figure 2-10 shows the analogous (to Figure 2-9) MOE values that result from the consideration of the 0.1 ng m^{-3} threshold. Overall, behavior similar to that described for the 0.01 ng m^{-3} threshold-based results is observed. The biggest difference associated with the MOE values at the two different thresholds can be seen by comparing the 12-hour RTW results. Whereas at the 0.01 ng m^{-3} threshold, under-predictions were initially indicated (Figure 2-9c and 2-9d) for SCIPUFF and ARAC, at the 0.1 ng m^{-3} threshold, initial SCIPUFF predictions suggest a slight over-prediction (Figure 2-10c) and ARAC shows little bias at any time period (Figure 2-10d).

When judging model predictive performance using the MOE based on the 0.01 ng m^{-3} threshold (as in Figure 2-9), one of two time-dependent behaviors is observed among several of the models. For some models, an initial under-prediction of the number of locations that exceed the threshold is followed by a “correction” that leads to about the right number of locations predicted above the threshold (movement to the 45-degree diagonal), followed finally, by degradation that suggests a general missing of the locations at which the threshold is exceeded (as described above). For other models, an initial over-prediction of the number of locations that exceed the threshold is followed by a “correction” that leads to about the right number of locations predicted above the threshold (movement to the 45-degree diagonal), followed again, by degradation that suggests a general missing of the locations at which the threshold is exceeded. Figure 2-11 illustrates these behaviors for 12-hour RTW threshold-based (0.01 ng m^{-3}) MOE values. The three plots on the left (108, 111, and 211) illustrate the first behavior described above (initial under-prediction, followed by relative correction, followed by missed locations) and the three plots on the right (115, 123, 207) illustrate the second behavior described above (initial over-prediction, followed by relative correction, followed by missed locations).

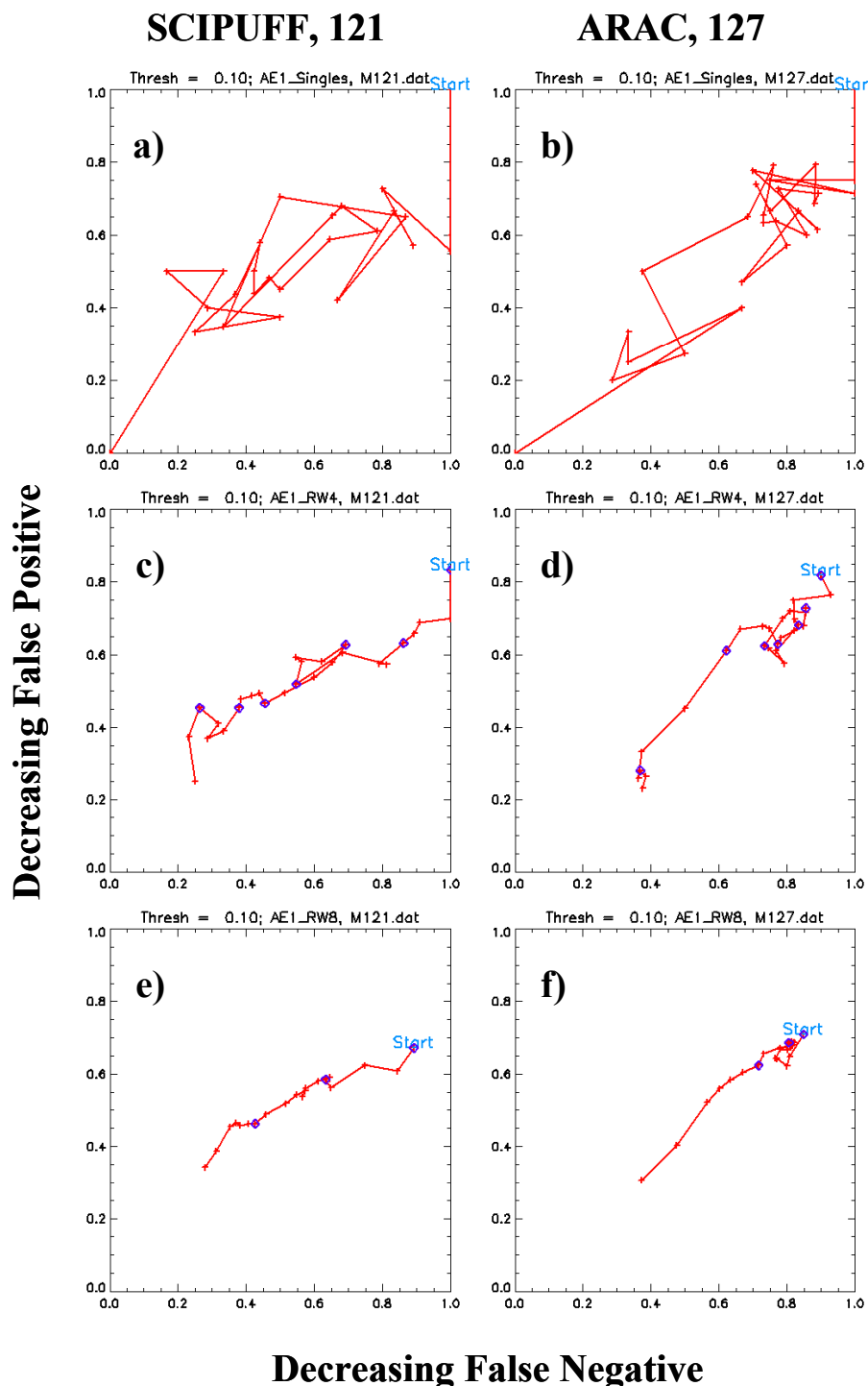


Figure 2-10. MOE Values Based on a Threshold Concentration of 0.1 ng m^{-3} for Model Predictions of *ETEX*. a) Values for SCIPUFF for the 30 Consecutive 3-Hour Periods, b) Values for ARAC for the 30 Consecutive 3-Hour Periods, c) Values for SCIPUFF for 12-Hour RTW, d) Values for ARAC for 12-Hour RTW, e) Values for SCIPUFF for 24-Hour RTW, and f) Values for ARAC for 24-Hour RTW. For the 12-Hour RTW Plots, MOE Values Colored Blue Diamonds Correspond to the “Independent” Values That Occur Every 12 Hours. For the 24-Hour RTW Plots, MOE Values Colored Blue Diamonds Correspond to the “Independent” Values That Occur Every 24 Hours.

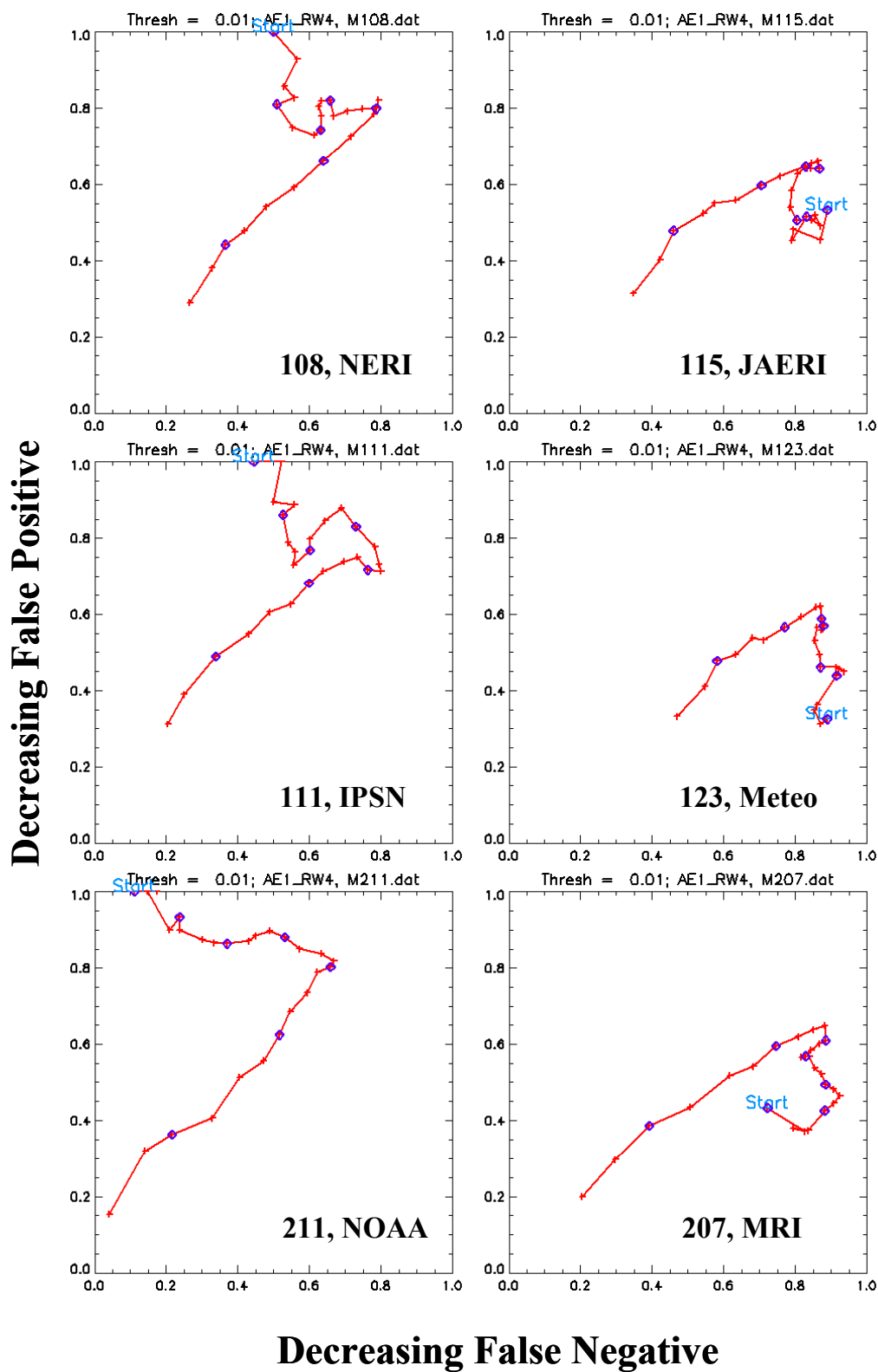


Figure 2-11. 12-Hour RTW MOE Values Based on a Threshold Concentration of 0.01 ng m^{-3} for Six Model Predictions of *ETEX*. MOE Values Colored Blue Diamonds Correspond to the “Independent” Values That Occur Every 12 Hours.

Figure 2-12 presents 12-hour RTW MOE results based on the 0.1 ng m^{-3} threshold. The time-dependent behaviors associated with models 211, 115, 123, and 207 are quite similar to those discussed at the 0.01 ng m^{-3} threshold level (Figure 2-11). However, for models 108 and 111, there is less initial under-prediction associated with the 0.1 ng m^{-3} threshold relative to the 0.01 ng m^{-3} threshold. This relative behavior is similar to that observed for SCIPUFF and ARAC.

Other time-dependent behaviors also can be seen. Figure 2-13 illustrates the time dependence of threshold-based MOE values for the two models that were ranked highest by OSF and RWFMS (1,1) – 202 and 105 (Table 2-1) – for the threshold-based (0.01 ng m^{-3}) MOE. Both 105 and 202 correspond to the CMC (Canadian) model (Table 1-1). For these model predictions the 0.01 ng m^{-3} threshold-based MOE values do not appear to degrade with time as much as the other models. Also, the time-dependent MOE values shown in Figure 2-13 cluster about the 45-degree diagonal, indicating that about the right number of locations were being predicted to have exceeded the threshold (i.e., neither an over- nor under-prediction).

Figure 2-14, analogous to Figure 2-13, shows 12-hour and 24-hour RTW MOE values for the two highest ranked (OSF and RWFMS (1,1) – Table 2-2) model predictions – 208 and 128 – based on a 0.1 ng m^{-3} threshold. Both 128 and 208 correspond to the SMHI (Swedish) model (Table 1-1). The time-dependent MOE values (for a 0.1 ng m^{-3} threshold) shown in Figure 2-14 cluster about the 45-degree diagonal, indicating that about the right number of locations were being predicted to have exceeded the threshold (i.e., neither an over- nor under-prediction). This is identical to the result observed for the 0.01 ng m^{-3} time-dependent MOE values. However, unlike the 0.01 ng m^{-3} threshold-based MOE values, the 0.1 ng m^{-3} threshold-based MOE show degradation at the longest time periods. This degradation in performance is associated with missing the locations (e.g., direction), even though the overall number of locations at which the threshold is exceeded is predicted about right (no bias). Also, comparison of the plots shown in Figure 2-14 to those of Figure 2-12 indicates that for the “highest ranked” predictions, 128 and 208, the start of degradation in performance as measured by the time-dependent MOE is delayed by several hours (about 12) relative to the nominal model predictions (Figure 2-12).

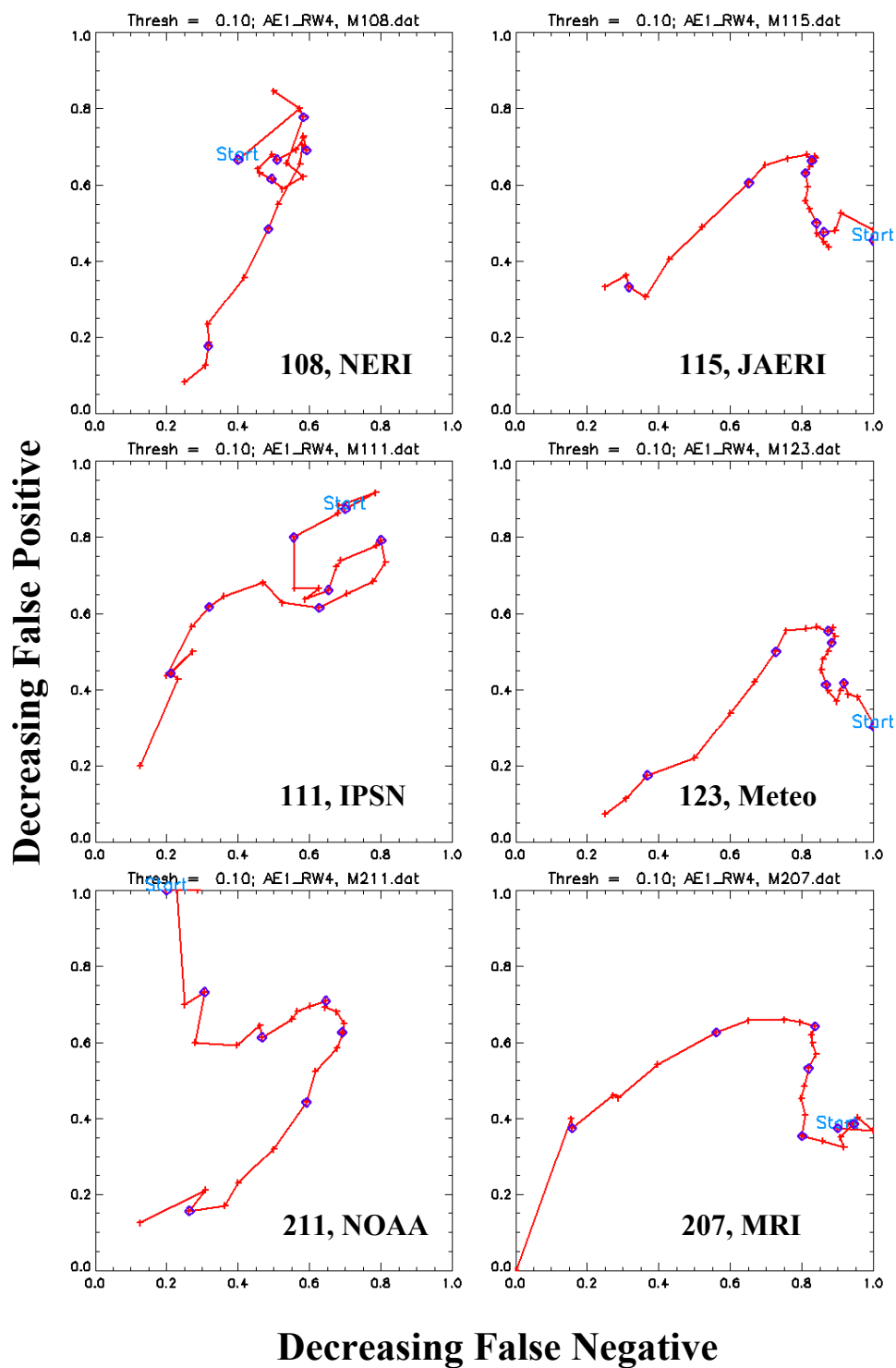


Figure 2-12. 12-Hour RTW MOE Values Based on a Threshold Concentration of 0.1 ng m^{-3} for Six Model Predictions of *ETEX*. MOE Values Colored Blue Diamonds Correspond to the “Independent” Values That Occur Every 12 Hours.

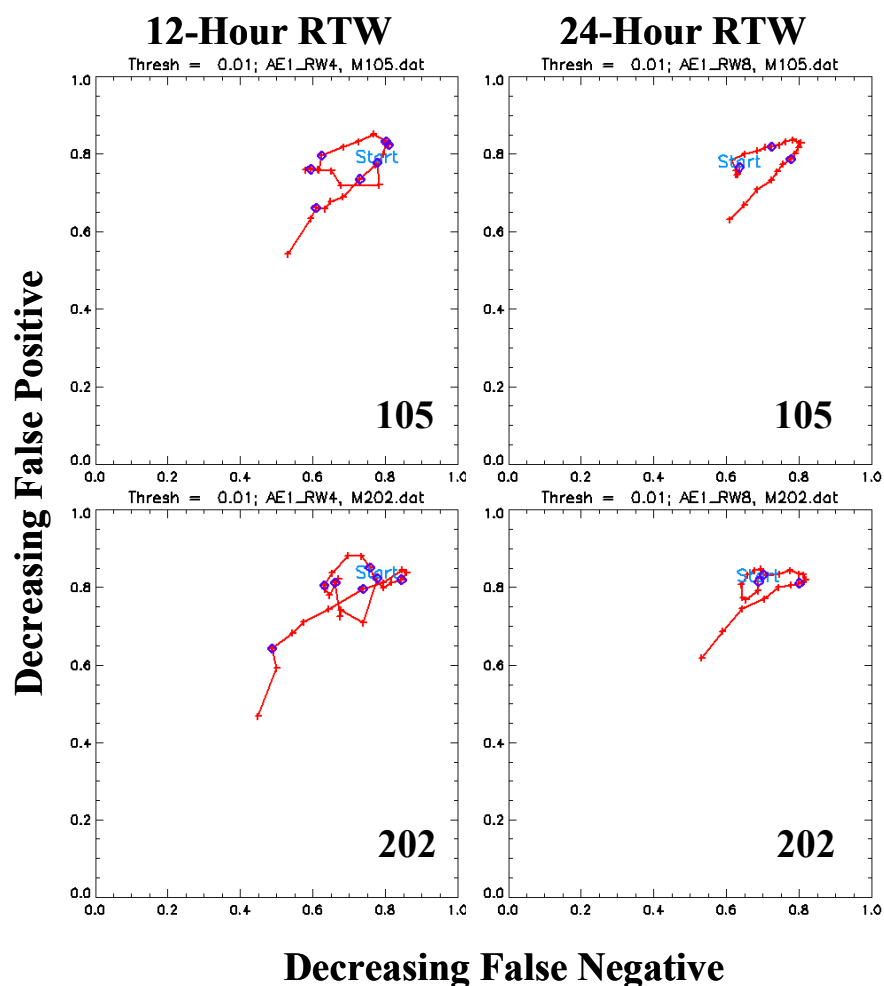


Figure 2-13. 12-Hour and 24-Hour RTW MOE Values Based on a Threshold Concentration of 0.01 ng m⁻³ for Two “Highly-Ranked” Model Predictions of *ETEX*. “105” and “202” Both Correspond to the CMC Model (Table 1-1). For the 12-Hour RTW MOE Values, Colored Blue Diamonds Correspond to the “Independent” Values That Occur Every 12 Hours. For the 24-Hour RTW MOE Values, Colored Blue Diamonds Correspond to the “Independent” Values That Occur Every 24 Hours.

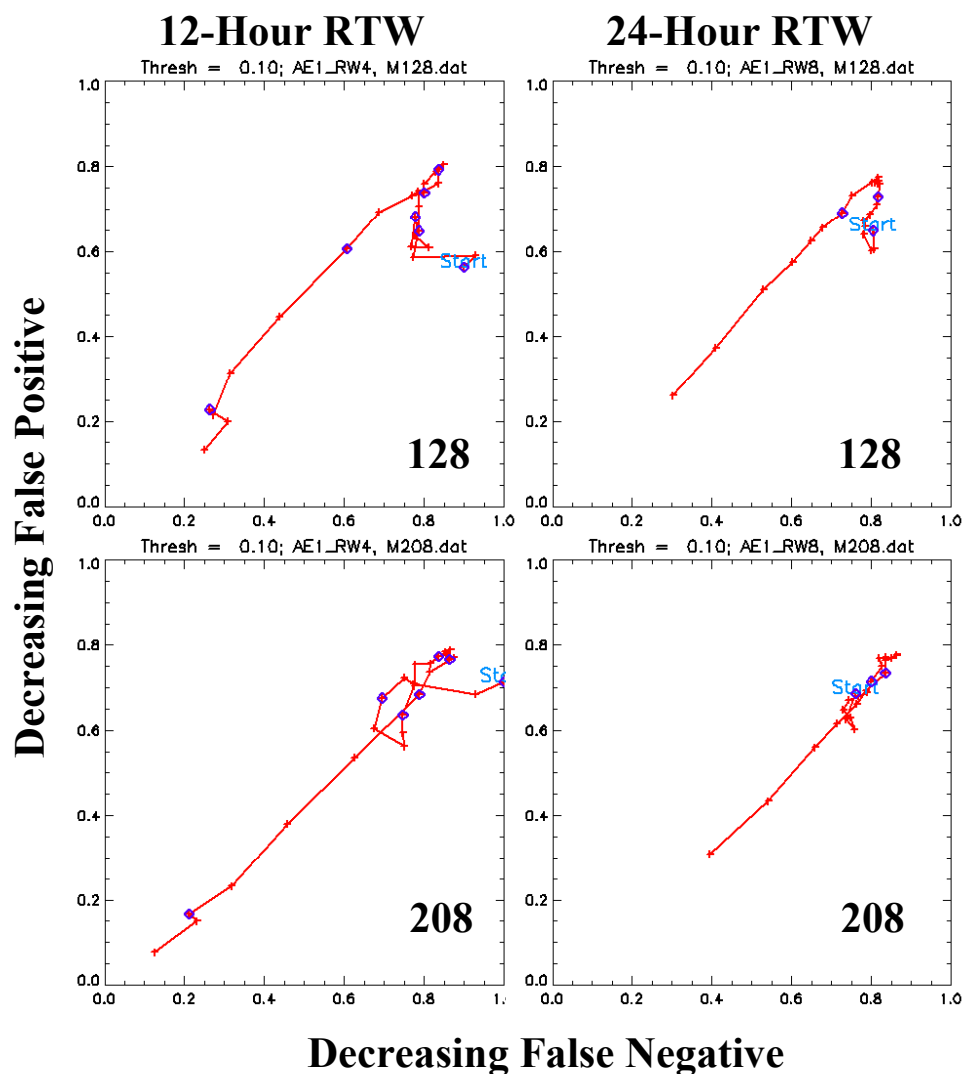


Figure 2-14. 12-Hour and 24-Hour RTW MOE Values Based on a Threshold Concentration of 0.1 ng m^{-3} for Two “Highly-Ranked” Model Predictions of *ETEX*. “128” and “208” Both Correspond to the SMHI Model (Table 1-1). For the 12-Hour RTW MOE Values, Colored Blue Diamonds Correspond to the “Independent” Values That Occur Every 12 Hours. For the 24-Hour RTW MOE Values, Colored Blue Diamonds Correspond to the “Independent” Values That Occur Every 24 Hours.

Time-dependent summed concentration-based MOE values were also examined. Figure 2-15 shows an interesting result. First, models 203 and 107 (both DWD, German) are ranked “1” and “2” in terms of absolute fractional bias (Table 2-4). Figure 2-15 shows the 24-hour RTW MOE values for these two sets of predictions. Although both sets of predictions achieved similar highly ranked performance, the time-dependent MOE values indicate substantially different behavior. Model 107, that included the ECMWF as the meteorological input, predicts the amount of material about right (MOE values are relatively near the 45-degree diagonal) over the entire 90-hour period. On the other hand,

model 203, that included modeler-selected meteorological input, resulted in an initial over-prediction followed by a later under-prediction. In terms of the absolute fractional bias scoring function, ABS(FB), these over- and under-predictions “cancelled” each other a bit and led to relatively good average performance in terms of FB.

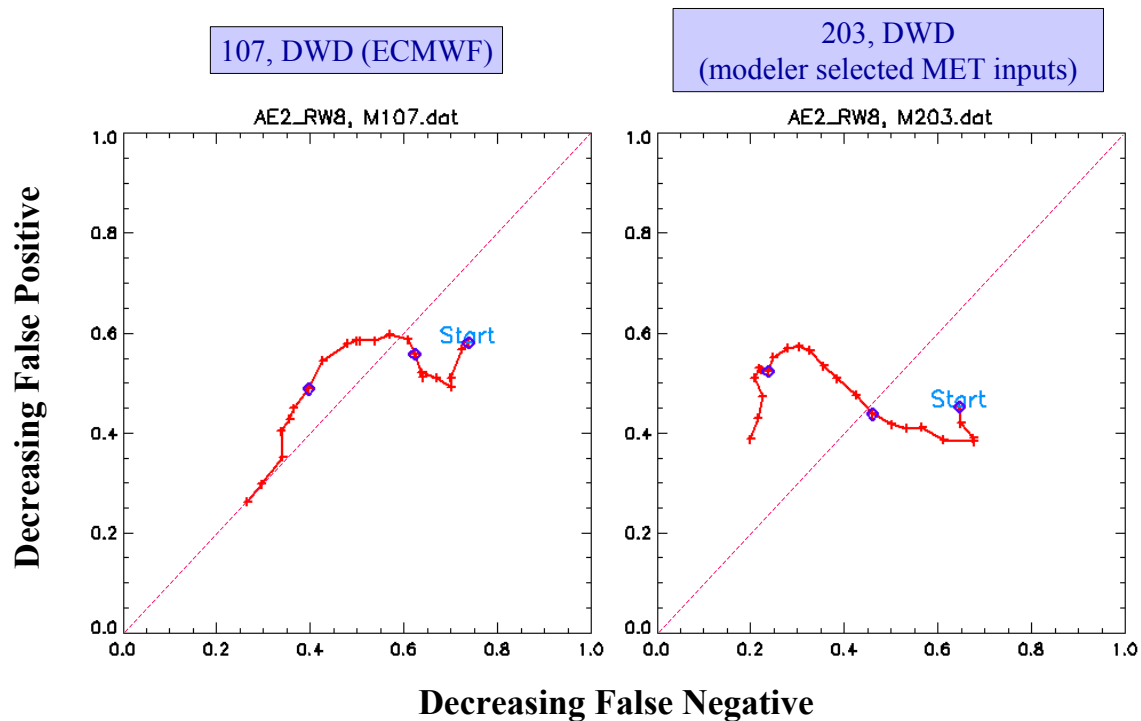


Figure 2-15. 24-Hour RTW Summed Concentration MOE Values for the Top Two ABS(FB) Ranked Model Predictions of *ETEX*. “203” and “107” Both Correspond to the DWD Model (Table 1-1). For the 24-Hour RTW MOE Values, Colored Blue Diamonds Correspond to the “Independent” Values That Occur Every 24 Hours.

D. PLANNED FUTURE STUDIES INVOLVING THE MOE AND *ETEX*

This study has demonstrated the usage of the user-oriented two-dimensional MOE to evaluate 46 model predictions of *ETEX*. Using a few scoring functions that could be identified with notional user requirements, these 46 models could be ranked in terms of the desired performance as specified by the scoring function. We also examined the sensitivity of MOE values to any single sampler location and found that evaluations of a few models’ performance was greatly affected by a single sampler location close to the release point. Finally, the usage of the MOE to explore the time-dependence of model performance was briefly introduced and described.

This study is intended to be a base upon which to build for future studies involving *ETEX*. First, we intend to create MOE values based on actual areas (e.g.,

square kilometers). Recall that no area interpolation was used in the present study. An important part of this next effort will be to explore and understand potential sensitivities associated with interpolation given the underlying non-uniform sampler space across Europe. Given area-based MOE values, one can then include European population distributions and notional effects-levels of interest to place the MOE in its ultimate context – fraction of the population falsely warned and fraction of the population inadvertently exposed. At this point the 46 models can be re-ranked given this more operational context. We also plan to create new HPAC (SCIPUFF) predictions of *ETEX*, to include probabilistic outputs, and evaluate the resulting predictions in terms of the user-oriented two-dimensional MOE.

REFERENCES

- 2-1. Warner, S. and Platt, N., 2003: *Analyses in Support of Initial Validation of Urban HPAC: Comparisons to Urban 2000 Observations*, IDA Document D-2870N, 119 pp, August 2003. (Available via an e-mail request to Steve Warner at swarner@ida.org or a mail request to Steve Warner, Institute for Defense Analyses, 4850 Mark Center Drive, Alexandria, Virginia 22311-1882.)
- 2-2. Warner, S., Platt, N., and Heagy, J. F., 2004: "Comparisons of Transport and Dispersion Model Predictions of the *Urban 2000* Field Experiment," submitted, *J. Appl. Meteor.*
- 2-3. Mosca, S., Bianconi, R., Bellasio, R., Graziani, G., and Klug, W., 1998: *ATEMS II – Evaluation of Long-Range Dispersion Models Using Data of the 1st ETEX Release*, Joint Research Center, European Commission, Office of Official Publications of the European communities, L-2985 (CL-NA-17756-EN-C), Luxembourg, 1998.
- 2-4. Sykes, R. I., et al., 2000: *PC-SCIPUFF Version 1.3 Technical Documentation*, A.R.A.P Report No. 725, Titan Corporation, ARAP Group, December 2000, pages 221- 226.

APPENDIX A

ACRONYMS

APPENDIX A

ACRONYMS

2D	Two-dimensional
ABS	Absolute value
A _{OB}	Region Associated With the Observations
A _{FP}	False Positive Region
A _{FN}	False Negative Region
ANPA	National Agency for Environment (Italy)
A _{OV}	Region of Overlap
A _{PR}	Region Associated With the Prediction
ARA	Applied Research Associates
ARAC	Atmospheric Release Advisory Center
ARAP	Aeronautical Research Associates of Princeton
ATEMS	Atmospheric Transport Model Evaluation Study
ATP	Allied Tactical Publication
BMRC	Bureau of Meteorology Research Center (Australia)
C _{FN}	false negative coefficient
C _{FP}	false positive coefficient
CMC	Canadian Meteorology Centre
CNR	National Research Council (Italy)
C _o	observed concentrations
C _p	predicted concentrations
DNMI	Norwegian Meteorological Institute
DMI	Danish Meteorological Institute
d _{OSF}	distance to (1,1) (for objective scoring function)
DOE	Department of Energy
DTIC	Defense Technical Information Center
DTRA	Defense Threat Reduction Agency
DWD	German Weather Service
ECMWF	European Centre for Medium Range Weather Forecasts
EDF	France Electricity
ETEX	European Tracer Experiment
FB	Fractional Bias
FBFOM	Fractional Bias Figure of Merit

FMS	Figure of Merit in Space
FOA	Defense Research Establishment (Sweden)
FOM	Figure of Merit
FY	Fiscal Year
HPAC	Hazard Prediction and Assessment Capability
IDA	Institute for Defense Analyses
IDL	Interactive Data Language
IMP	Institute for Meteorology and Physics, University of Wien (Austria)
IMS	Swiss Meteorological Institute
IPSN	French Institute for Nuclear Protection and Safety
JAERI	Japan Atomic Research Institute
JMA	Japan Meteorological Agency
KMI	Royal Institute of Meteorology of Belgium
LLNL	Lawrence Livermore National Laboratory
Meteo	Meteo France
MetOff	Meteorological Office (United Kingdom)
MOE	Measure of Effectiveness
MRI	Meteorological Research Institute (Japan)
MSC-E	Meteorological Synthesizing Centre – East (Russia)
NARAC	National Atmospheric Release Advisory Center
NAD	Normalized Absolute Difference
NERI	National Environment Research Institute / Risoe National Laboratory/ University of Cologne (Germany / Denmark)
ng m ⁻³	nanograms per cubic meter
NIMH-BG	National Institute of Meteorology and Hydrology (Bulgaria)
NIMH-BG	National Institute of Meteorology and Hydrology (Romania)
NMSE	Normalized Mean Square Error
NOAA	National Oceanic and Atmospheric Administration
OLAD	Over-Land Along-Wind Dispersion
OSF	Objective Scoring Function
PMCH	Perfluoro-methyl-cylcohexane
Ref.	Reference
RTW	Running Time Window
RWFMS	Risk-Weighted Figure of Merit in Space

SAIC	Science Applications International Corporation
SBIR	Small Business and Innovative Research
SCIPUFF	Second-Order Closure Integrated Puff
SMHI	Swedish Meteorological and Hydrological Institute
SRS	Westinghouse Savannah River Laboratory
T&D	Transport and Dispersion
UTC	Universal Time Coordinated
V&V	Verification and Validation
VV&A	Verification, Validation, and Accreditation
WMD	Weapons of Mass Destruction

APPENDIX B

TASK ORDER EXTRACT

APPENDIX B

TASK ORDER EXTRACT

DC-9-1797

TITLE: Support for DTRA and LLNL in the Validation Analysis of Hazardous Material Transport and Dispersion Prediction Models

This task order is for work to be performed by the Institute for Defense Analyses (IDA) under Contract DASW01-98-C-0067 and DASW01-02-C-0012 for the Defense Threat Reduction Agency (DTRA).

1. BACKGROUND:

The Hazard Prediction and Assessment Capability (HPAC) is a suite of codes that predicts the effects of hazardous material releases into the atmosphere and their impact on civilian and military populations. The software can use integrated source terms, high-resolution weather forecasts, and particulate transport models to predict hazard areas produced by battlefield or terrorist use of weapons of mass destruction (WMD), by conventional counterforce attacks against WMD facilities, or by military and industrial accidents.

The DTRA Verification and Validation (V&V) Program represents ongoing activities performed in parallel with development of all predictive codes in support of HPAC. One element of V&V is to perform code-on-code comparisons. In this strategy, each code receives the same input. In this manner, differences in the output predictions can lead to the identification of software bugs, or help to assess technical strengths and weaknesses of component algorithms within each code. In addition, a certain amount of credibility for both models is achieved when their predictions agree. When the inputs are simple, such as for fixed winds and simple terrain, the predictions tend to be dominated by the dispersion algorithms. Comparisons at this level of complexity are important to establish fundamental dispersion algorithm veracity, and to help discover software bugs. As more complex terrain and weather is included as input, the number of physical processes responsible for transport and dispersion increases and the predictions become the result of many interdependent algorithm calculations.

Code-on-code comparisons will be performed using the DTRA code HPAC, the Lawrence Livermore National Laboratory (LLNL) code National Atmospheric Release Advisory Capability (NARAC), and, possibly, other government-developed codes. These codes represent major national investments in transport and dispersion modeling within their respective applications. The comparisons will provide information from

which to validate the HPAC and NARAC models (and perhaps others), as well as provide an opportunity to advance both technologies. The code comparisons will include short-, medium-, and long-range transport distances. Complex terrain and weather will also be included.

It is very difficult to separate meteorological uncertainty from the transport and dispersion model accuracy when comparing predictions to field-trial validation quality or real-world data. The validation challenge is to assess whether a model performs well over different field trials, and ultimately reflects real-world phenomena. Some codes perform better under certain conditions and specific scenarios. Hazard prediction models are generally developed for a range of user communities and applications. Each user community has a different set of requirements. Thus, the corresponding hazard models tend to be optimized for specific applications. The process of accrediting a model is always couched in terms of the end-user requirements.

Various figures-of-merit (FOM) are used to express model performance relative to observed data. Most FOMs tend to use manifestations of a ratio (geometric or arithmetic) between the predicted and observed quantities. The compared quantities are usually peak, plume-centerline, and off-axis concentration or dosage, as well as crosswind and along-wind spread and area coverage. Other FOMs may include the second-moment of the dosage and concentration values at a sampler location. All these FOMs are reasonable measures, but none of them explicitly expresses application-oriented performance. A “yardstick” is needed that measures application-oriented model performance. The scale on this yardstick would clearly and directly relate to the specific user’s concerns and needs. The pursuit of this “accreditation” performance measure is a continuing initiative at DTRA.

2. OBJECTIVE:

IDA will conduct independent analysis and special studies associated with verification and validation of the suite of models associated with the Hazard Assessment and Prediction Capability. IDA will support development of user-oriented performance measures of effectiveness (MOE) using validation quality field trial data sets; coordinate scenario definition and arbitration for code-on-code V&V activities; and assist DTRA and the Department of Energy in identifying the V&V parameter space associated with various hazard assessment and collateral effects communities.

The objectives of verification and validation analysis and coordination are: (1) to ensure that a consistent analysis approach is used when comparing model predictions, and assist DTRA in the implementation of code-on-code analysis, comparisons, and interpretation; and (2) to define and further develop measures of effectiveness in terms of user-specific objectives and applications.

The scope of this effort may be expanded to other programs as directed by DTRA.

3. STATEMENT OF WORK:

As required by DTRA technical representatives, IDA will perform the following tasks:

- a. Advanced User-Oriented Measure of Effectiveness (MOE) Development

IDA will conduct model prediction to field trial observation comparisons using a novel user-oriented MOE. Mean value and probabilistic prediction outputs (e.g., from HPAC) will be examined and relative performance will be described.

For fiscal year 2003, comparisons of model predictions to field trial data at mid-range (tens of km, e.g., OLAD)) and long range (hundreds of km, e.g., *ETEX*) will be conducted and reported. Model prediction comparisons (e.g., NARAC and HPAC) via the MOE will be conducted for those data sets. In addition, the inclusion of different predictive weather inputs (“weather experts”) may be considered within the framework of model validation/accreditation.

b. Comparisons of DTRA-Identified Urban and Building Interior T&D Models

For FY 03, IDA will begin a substantial effort to compare the predictions of DTRA-identified urban and building interior T&D codes to field trial data. IDA will also continue to extend the application of the user-oriented MOE to building interior and urban models of hazardous material transport and dispersion.

Various sampler weighting and interpolation schemes that can be applied to building interior, urban transport, and longer-range data sets will be explored, compared, and contrasted.

c. Communication: Using the MOE for Model Accreditation

IDA will focus particular effort on the communication, via various methods, of the value, usage, and technical merits of the new validation and accreditation MOE. Technical and operator review and feedback will be sought and considered.

(1) For FY 2003, IDA will continue the development of a “demonstration” accreditation. This effort will require the identification of a potential user and specific application. For this user(s) and application(s), IDA will focus on extracting a sense for what are the acceptable user requirements (i.e., risk tolerance). These requirements will differ among potential user groups (military targeting, passive CB defense, civilian first responders, military versus civilian population human effects, etc.). Similarly, previously described lethality/effects filters will be used to interpret MOE results and reviewed with potential users. The goal of the above effort is to demonstrate the “end-to-end” accreditation of a model usage (e.g., a particular HPAC probabilistic output) for a specific application and user (i.e., agreed to/acceptable risk tolerance). The chosen application and user should correspond to an actual situation (i.e., not simply represent a notional scenario).

(2) Appropriate comparisons of model output (HPAC) and ATP-45 hazard areas are the goal of an additional FY 2003 effort. Notional scenarios for comparison will be chosen so as to help elucidate fundamental differences between ATP-45 and HPAC “predictions.” With this effort, we hope to identify situations in which the model represents and operational improvement over ATP-45 (e.g., in terms of the user-oriented MOE).

(3) IDA will communicate, via conference papers and/or posters, working group discussions, IDA papers, and peer-reviewed journal articles, the more

important applications of the MOE and any progress toward the creation of a “demonstration” accreditation.

d. Comparisons to Other T&D Models

As required, IDA will continue to provide coordination for model comparisons (that is, HPAC comparisons, to other models and field trial data). For example: (1) IDA will support the selection of longer-range field trial data for future model comparisons, and (2) working with the DOE’s Los Alamos National Laboratory, IDA may consider DOE field trial data sets to validate the new fire and explosion source terms that are being introduced into HPAC. Additionally, IDA may conduct studies of specific HPAC features and algorithms where issues arise (e.g., aero-breakup modeling algorithms or weather assimilation features) or are identified by the sponsor. As in the past, IDA will also coordinate the analyses and reporting of such comparisons.

Finally, IDA may use their MOE’s to review and provide comment on an inverse, adjoint plume model in development under an SBIR to Aerodyne, Inc. The initial goal of this Phase II SBIR is to provide location and yield information on nuclear events from data collected by a worldwide network of monitoring stations.

4. CORE STATEMENT:

This research is consistent with IDA’s mission in that it will support specific analytical requirements of the sponsor and will assist the sponsor with planning efforts. Accomplishment of this task order requires an organization with experience in operationally oriented issues from a joint and combined perspective, which IDA, a Federally Funded Research and Development Center, is able to provide. It draws upon IDA’s core competencies in Systems Evaluations and Operational Test and Evaluation. Performance of this task order will benefit from and contribute to the long-term continuity of IDA’s research program.

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
<small>Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.</small>					
1. REPORT DATE (DD-MM-YYYY) 01-11-2003		2. REPORT TYPE Final		3. DATES COVERED (From - To) Oct. 2002 - Oct. 2003	
4. TITLE AND SUBTITLE Application of User-Oriented MOE to Transport and Dispersion Model Predictions of the European Tracer Experiment				5a. CONTRACT NUMBER DASW01-98-C-0067 / DASW01-02-C-0012	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Steve Warner, Nathan Platt, James Heagy				5d. PROJECT NUMBER	
				5e. TASK NUMBER DC-9-1797	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Institute for Defense Analyses 4850 Mark Center Drive Alexandria, VA 22311-1882				8. PERFORMING ORGANIZATION REPORT NUMBER IDA Paper P-3829	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Defense Threat Reduction Agency 6801 Telegraph Road, Room 264 Alexandria, VA 22310-3398				10. SPONSOR/MONITOR'S ACRONYM(S) DTRA	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution unlimited. Defense Threat Reduction Agency, Ft. Belvoir, Virginia, 2 December 2003.					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT In October 1994, the tracer gas perfluoro-methyl-cyclohexane (PMCH) was released over a 12-hour period from a location in northwestern France and tracked at 168 sampling locations in 17 countries across Europe (100s of kilometers). This release, known as the European Tracer Experiment (ETEX), resulted in the collection of a wealth of data. IDA has obtained the predictions of 46 transport and dispersion models from 17 countries from the Joint Research Centre, European Commission as well as the PMCH sampling data. This paper describes the extension of the previously described user-oriented measure of effectiveness (MOE) methodology to evaluate the predictions of the 46 models against the long-range ETEX observations. This paper develops the methodological protocols to compare model predictions of ETEX using the MOE and to score and rank model performance by a variety of notional user criteria. The development of these notional user criteria is also described in this paper. In addition, the sensitivity of MOE estimates to any single sampler location is examined and the use of the MOE to explore the time-dependence of model performance is also described.					
15. SUBJECT TERMS model validation; hazardous material transport and dispersion; HPAC; ETEX; measure of effectiveness					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT Unlimited	18. NUMBER OF PAGES 86	19a. NAME OF RESPONSIBLE PERSON Mr. Richard Fry
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified			19b. TELEPHONE NUMBER (include area code) 703-325-6760